Artificial Neural Networks in Greenhouse Modelling

Two modelling applications in horticulture

Dissertation

zur Erlangung des akademischen Grades
doctor rerum agriculturarum
(Dr. rer. agrar.)

eingereicht an der Lebenswissenschaftliche Fakultät
der Humboldt-Universität zu Berlin

von

Luis Carlos Miranda Trujillo

Präsidentin der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät

Prof. Dr. Bernhard Grimm

Gutachter:

1. Prof. Dr. Uwe Schmidt
2. Dr. Bruno Lara
3. Prof. Dr. Verena Hafner

Tag der mündlichen Prüfung: 28.07.2018

Artificial Neural Networks in Greenhouse Modelling

Two modelling applications in horticulture

Luis Carlos Miranda Trujillo

SUMMARY

One facet of the current developments in precision horticulture is the highly technified production under cover. The intensive production in modern greenhouses heavily relies on instrumentation and control techniques to automate many tasks. Among these techniques are control strategies, which can also include some methods developed within the field of Artificial Intelligence. This document presents research on Artificial Neural Networks (ANN), a technique derived from Artificial Intelligence, and aims to shed light on their applicability in greenhouse vegetable production. In particular, this work focuses on the suitability of ANN-based models for greenhouse environmental control. To this end, two models were built: A short-term climate prediction model (air temperature and relative humidity in time scale of minutes), and a model of the plant response to the climate, the latter regarding phytometric measurements of leaf temperature, transpiration rate and photosynthesis rate. A dataset comprising three years of tomato cultivation was used to build and test the models. It was found that this kind of models is very sensitive to the fine-tuning of the metaparameters and that they can produce different results even with the same architecture. Nevertheless, it was shown that ANN are useful to simulate complex biological signals and to estimate future microclimate trends. Furthermore, two connection schemes are proposed to assemble several models in order to generate more complex simulations, like long-term prediction chains and photosynthesis forecasts. It was concluded that ANN could be used in greenhouse automation systems as part of the control strategy, as they are robust and can cope with the complexity of the system. However, a number of problems and difficulties are pointed out, including the importance of the architecture, the need for large datasets to build the models and problems arising from different time constants in the whole greenhouse system.

ZUSAMMENFASSUNG

Moderne Präzisionsgartenbaulicheproduktion schließt hoch technifizierte Gewächshäuser, deren Einsatz in großem Maße von der Qualität der Sensorik- und Regelungstechnik abhängt, mit ein. Zu den Regelungsstrategien gehören unter anderem Methoden der Künstlichen Intelligenz, wie z.B. Künstliche Neuronale Netze (KNN, aus dem Englischen). Die vorliegende Arbeit befasst sich mit der Eignung KNN-basierter Modelle als Bauelemente von Klimaregelungstrategien in Gewächshäusern. Es werden zwei Modelle vorgestellt: Ein Modell zur kurzzeitigen Voraussage des Gewächshausklimas (Lufttemperatur und relative Feuchtigkeit, in Minuten-Zeiträumen), und Modell zur Einschätzung von phytometrischen Signalen (Blatttemperatur, Transpirationsrate und Photosyntheserate). Eine Datenbank, die drei Kulturjahre umfasste (Kultur: Tomato), wurde zur Modellbildung bzw. -test benutzt. Es wurde festgestellt, dass die ANN-basierte Modelle sehr stark auf die Auswahl der Metaparameter und Netzarchitektur reagieren, und dass sie auch mit derselben Architektur verschiedene Kalkulationsergebnisse liefern können. Nichtsdestotrotz, hat sich diese Art von Modellen als geeignet zur Einschätzung komplexer Pflanzensignalen sowie zur Mikroklimavoraussage erwiesen. Zwei zusätzliche Möglichkeiten zur Erstellung von komplexen Simulationen sind in der Arbeit enthalten, und zwar zur Klimavoraussage in längerer Perioden und zur Voraussage der Photosyntheserate. Die Arbeit kommt zum Ergebnis, dass die Verwendung von KNN-Modellen für neue Gewächshaussteuerungstrategien geeignet ist, da sie robust sind und mit der Systemskomplexität gut zurechtkommen. Allerdings muss beachtet werden, dass Probleme und Schwierigkeiten auftreten können. Diese Arbeit weist auf die Relevanz der Netzarchitektur, die erforderlichen großen Datenmengen zur Modellbildung und Probleme mit verschiedenen Zeitkonstanten im Gewächshaus hin.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial Neural Networks
ANOVA	Analysis of variance
ARIMA	Autoregressive Integrated Moving Average Model
ARMA	Autoregressive Moving Average Model
ARX	Autoregressive Model with Exogenous Inputs
$both$	Model built with data from the collector and reference greenhouses
CM	Climate prediction model
col	Model built with data from the collector greenhouse
DFS	Days after the sowing date
ET_0	Reference Evapotranspiration
LAI	Leaf Area Index
LSTM	Long short-term memory neural network
LTP	Long-term prediction
M_{min}	Model built with a minimum set of inputs
M_{max}	Model built with all the inputs available
MAE	Mean absolute error
MLP	Multilayer perceptron neural network
MSE	Mean square error
NIC	Network Information Criterion
OSP	One-step prediction
$P-M$	Penman & Monteith evapotranspiration equation

PM	Model of phytometric data
PMa	Leaf temperature model
PMb	Transpiration rate model
PMc	Photosynthesis rate model
ref	Model built with data from the reference greenhouse
$rmse$	Root-mean-square error
VPD	Vapour pressure deficit of the air

1

INTRODUCTION

1.1 SCOPE AND MOTIVATION

Greenhouse horticulture, the intensive production of vegetables and ornamental plants under a transparent covering, is a many-coloured technology. Greenhouse technology comprises the *Parral*-type greenhouses in Almería, Spain, as well as the *Venlo*-type greenhouses in The Netherlands. Greenhouses range from the *Chinese Solar Greenhouse* to low- and high-tech multitunnel plastic houses elsewhere. The materials and structures can be different: wooden structures are found in the *Guadua* greenhouses in Colombia and the *Parral* greenhouses in Spain, steel in many other places. While the most common covering materials are plastic films and glass panes, there are differences in the installation, isolation, suplemental materials (for example, thermal screens or shadows), or even whitening in summertime.

But all greenhouses aim at protecting the crop. They build a physical barrier to protect the plants from pests, and animals in general, from weeds and dust, but also from adverse climatic conditions. Because the isolated volume created by the greenhouse covering has an important characteristic: it makes it possible to tune and control the plants' microclimate (Takakura and Fang, 2002).

However, this protection is an expensive one, and the growers try to maximise the production, in order to make the best out of their investment. Among the tools available for that effect is the automatic control of the climate and irrigation, including the use of mathematical knowledge regarding the plants and their environment. Matematical models and descriptions play a role in the quality of greenhouse automation, as they give a basis for the control actions to be taken. For example, previous experience (both scientific and practical) has shown that a crop often increases its water requirements under strong solar radiation, a knowledge useful for irrigation design.

As was mentioned before, greenhouses can be very different in construction. Additionally, the plants play an active role in shaping the evironment they grow in. These two factors make it difficult to develop mathematical models that are accurate enough for automation purposes. Indeed, as early as the speaking plant approach was introduced (Udink ten Cate et al., 1978), the greenhouse control strategies have looked for methods that can cope with the complexity of the whole technical-biological system.

This thesis deals with Artificial Neural Networks (ANN), as previous research work (Section 2.2.2) has shown that their capability to model complex systems can be useful to overcome the aforementioned challenges. More concretely, it deals with two parts of the greenhouse biophysical system: the physical conditions of the air and the plant processes taking place under those conditions. In the first case, the air conditions are described by the temperature and relative humidity. In the second, the plant processes (phytometric signals) of interest are the leaf temperature, the transpiration rate and the photosynthesis rate. Throughout the thesis, the focus is laid on the potential use of the ANN as constituting blocks for greenhouse control systems. However, control itself falls out of the scope of this work.

As redundant as it may sound, data-based modelling techniques (like ANN) rely on the disponibility of big amounts of data to construct the models. The worldwide trend to automatically store data (be it by automation computers or data loggers), as well as internet access and data-sharing possibilities, both in research and production greenhouses, allow to extend the use of these techniques. The use of these possibilities could mean for the horticultural sector a technological advancement similar to the one we witness in other branchs of the industry and society.

1.2 OBJECTIVES

In order to contribute to the further development of the greenhouse engineering field described previously in this chapter, the present dissertation addresses the following research questions:

- How can ANN be used to predict the climate inside a greenhouse?

- Can the predictions be improved by incorporating previous data records?
- Which information is important for a neural model of transpiration? Which information is important for a neural model of photosynthesis?
- What importance has the weather outside the greenhouse?
- How sensible are the simulations to a neural model's architecture?
- How can models be coupled to achieve more complex simulations?

Considering the fact that several research questions present themselves as transversal to different modelling tasks, the main objectives of this thesis are:

1. To implement predictive neural models of the greenhouse air temperature and relative humidity in a time scale of minutes.
2. To implement neural models of the following phytometric signals:
    - Leaf temperature
    - Transpiration rate
    - Photosynthesis rate

1.3 APPROACH

The two main objectives defined previously set the lines that define this complete thesis: This work aims at shedding light on the research questions by directly tackling these two design tasks. This approach to problem solving implies that the design and implementation processes bring answers and insights along with the models themselves. Therefore, the topics *climate prediction* and *pytometric signals* build the basic structure of Chapters 3 to 5.

For the design and implementation, special attention was put on the applicability and modularity of the models, keeping an eye at the intended application for automation purposes. This means that not the causal relationships between variables was on the focus, as would be the case

with mechanistic or first-principles models (Section 2.1). Rather, priority was given to their potential to enhance the performance of the models. This does not mean, however, that a purely pragmatic approach was adopted: the empirical and scientific horticultural expertise was considered all along the design and evaluation of the models.

The latter defines a second pair of recurring topics in this work: the purely technical data crafting using ANN on the one hand and the botanical science and horticultural practice in the other. Both topics set constrains and working lines that needed to be taken into account simultaneously at all times.

In regard to the analysis of the models, both during design and test, a dual approach, using numerical measures and graphical descriptions, was adopted. Yet, while the models were quantitatively compared between themselves and with those reported in the literature, more emphasis was given to their qualitative performance under situations of interest. There are two main reasons for this emphasis. Firstly, model comparison is a delicate question if the modelling conditions, boundaries and techniques are not clearly defined and indeed comparable. This is not always the case with models reported in the literature, where different error measurements and different boundary conditions are often reported with no standard structure. Secondly, it was particularly interesting to search for technical issues that might require more care in the design process, as a means to build up recommendations and expertise for further research. Thus, problems, errors and pitfalls were explicitly looked for and examined with care. These conditions are seldom accounted for with simple numerical measures of error.

The experimental setup framed by the ZINEG project (Section 3.1.1) made it possible to address the issue of different facilities and their corresponding models, if only with two greenhouses. The theme was addressed by building three different neural models for climate prediction as well as for phytometric signals: one built with data from each greenhouse, and the third one with a mixed dataset. Even though these results should be taken with a grain of salt, due to the high similarity of the greenhouses and the growing conditions, they can give interesting insights towards models that can be used in several facilities.

1.4 CONTRIBUTION

This dissertation contributes to the area of greenhouse engineering by deepening the study of neural models and both their capacities and limitations when dealing with the greenhouse as a bio-technical system.

In particular, this thesis emphasises the model design process, directly addressing common misunderstandings related with neural models. This includes a clear assessment of the degree to which the neural models intrinsically differ in performance, due to the random weight initialisation, an element seldom found in published research in the field. It is also shown that neither an increase of model complexity (by addition of hidden nodes to the networks) nor the use of more input signals necessarily increase the model performance. The latter point is particularly discussed on the light of an apparent conflict with causal relationships found in mechanistic models, for example in the case of the solar radiation as motor force to both the air temperature (in case of the climate models) and the plant transpiration and photosynthesis (when dealing with models of phytometric signals). In the light of the mentioned variability of the model's results, it is shown in Section 4.1.1 and Section 4.2.1 that an error measure alone does not always suffice as model selection criterion, because variations in model architecture and inputs would often fall inside this random range. Therefore, in addition to statistical tests, a qualitative, descriptive approach based on selected examples was used to complement the model design and test. This descriptive approach allowed to identify errors in the models (particularly the role of the ventilation and screening) that would remain concealed behind a general measure of error.

In addition to the methodological aspects mentioned before, the following results are of particular novelty in the field. Firstly, the connection scheme developed for the joint simulation of the transpiration and photosynthesis rates (Figure 46 on page 103) builds on biological knowledge to enhance the neural simulations and offers an architecture that can be useful in cases with limited instrumentation. Secondly, the topic was further elaborated to propose a connection scheme (Figure 49 on page 107) that links the models of climate and phytometric signals and thus profit from the calculation of the expected response of the plant canopy to current control actions taken by the automation system. Such architectures can be included in plant-based predictive controllers. Lastly, this is to

the best of our knowledge the first work to report experiments regarding cross-tests of neural models and greenhouse facilities, thus opening a line of research that we are convinced is of capital importance to the practical use of ANN in automation computers used in production.

1.5 STRUCTURE

The rest of this thesis is organised as follows.

Chapter 2 gives an overview of the research work that shapes the frame of reference for the experiments presented in further chapters. It is divided in two parts. First, Section 2.1 presents the use of mathematical models in greenhouse horticulture in two big areas: models of the greenhouse climate (Section 2.1.1) and models of the processes by the plants growing inside (Section 2.1.2). This section also sets the leitmotiv of the whole thesis, since further chapters are organised following the core division in *climate models* and *models of phytometric signals*. The second part of Chapter 2 deals with ANN, providing the mathematical and technical background needed for their application (Section 2.2.1) as well as a survey of their research usage in greenhouses (Section 2.2.2) thus far.

Chapter 3 describes in detail the formal aspects of the experiments lying at the core of this thesis. In its first part (Section 3.1), it gives information about all the sources of data that were used in the modelling experiments, including a description of the greenhouses themselves, the crop and sensors. Sections 3.1.4 and 3.1.5 present models and calculations that were carried out without sensors, and provided additional information to the neural network models. Section 3.2 focuses on the data manipulation carried out before the models could be trained, thus converting the raw data into a dataset. The same section deals with the separation of datasets for training, validation and test of the models (a concept that is explained in Section 2.2.1). Lastly, Section 3.3 deals with the model construction, including network architecture, concrete inputs and outputs for each model considered and the actual software implementation. Section 3.3 also introduces a connection scheme for the three phytometric signals considered. This connection scheme (Figure 13, on

page 56) allowed to use values of leaf temperature and transpiration rate to improve the simulation of transpiration and photosynthesis.

The main results of the experiments are presented in Chapter 4, with the first part (Section 4.1) devoted to the climate models and the second (Section 4.2) to the models of phytometric signals. Section 4.1.1 and Section 4.2.1 deal with the model construction process, including the selection of inputs and the validation of different model candidates, including the amount of previous information useful for climate prediction. The results of the model tests, shown in Section 4.1.2 and Section 4.2.2, are in both cases given first in terms of the complete dataset, complemented by selected examples consisting of two-days simulation runs. Additionally, Section 4.3 proposes a connection scheme that can be used to link the climate prediction model with the model of phytometric signals. A candidate for this connection scheme can be seen on Figure 49, on page 107.

Chapter 5 puts the experimental results in context, according to published research, and aims at an explanation of the problems found. The chapter has the same structure as the previous one: Section 5.1 discusses the results of climate prediction, including the long-term prediction of the temperature and relative humidity. Section 5.2 deals with the models of leaf temperature, transpiration and photosynthesis. Section 5.3 analises the results of the coupled models and the long-term prediction of the phytometric signals.

Lastly, Chapter 6 summarises the main findings of the thesis, with an outlook of possible applications as well as suggestions for further development.

2

LITERATURE REVIEW

2.1 MATHEMATICAL GREENHOUSE MODELS

Mathematical models are simplified representations of the reality. They use the language of mathematics, in the form of sets of equations, to mirror some part of a real system that is of interest. The assemblage of the equations that ultimately consitute a mathematical model requires a trade-off between the theoretical foundations of (and assumptions about) the underlying structure and processes of the analysed system on the one hand, and the computational capabilities available to draw conclusions from complicated models on the other hand (Bender, 2012). This design process implies an analysis that helps to understand the problem to be modeled, as well as a permanent focus on the model objectives and desired applications.

Depending upon the objectives and intentions of the modelling, the model itself can take one of two general forms:

- *Mechanicistic* models (also called *explanatory*, *theoretical*, *first-principles* or *white-box* models) provide a degree of understanding or explanation of the phenomena being modelled, and include the modeller's ideas of how the system works.
- *Descriptive* models (also called *empirical*, *regression*, *statistical* or *black-box* models) aim principally to describe the responses of a system, wihtout making any attempt to explain the mechanisms driving the phenomena.

Mechanistic models seek to gain insights and to provide an objective explanation of the system's behaviour (Lentz, 1998). Descriptive models aim to predict the response of a system in order to manage it, in which case it is enough (and indeed necessary) to ensure a good agreement of the model output variables with the reality they represent (Berg

and Kuhlmann, 1993). Mechanistic models help with understanding the system more than descriptive models do, and therefore tend to be more research-oriented than application-oriented (Wallach et al., 2014; Thornley and France, 2007). Conversely, descriptive models can be suitable for use in on-line control (Lentz, 1998) and as a decision support tool (Jones, 2014), as they profit from short computing time and implicitly take into account all unknown effects not included in the model design (Marcelis et al., 1998).

This model classification, however, often fails to picture the fact that most models are built in a complex manner and thus include aspects of both approaches. For example, in the lowest abstraction level, all mechanistic models rely on some form of descriptive ones (Marcelis et al., 1998; Thornley and France, 2007). Likewise, all descriptive models reflect the understanding of the modeller and incorporate different degrees of explanatory variables (Linker and Seginer, 2004; Ljung, 2010).

It can therefore be argued that both model approaches tend to meet in the center, as the theory underlying mechanistic models strives to be corroborated through experimental measurements, and data-driven empirical models refine their results through incorporation of theoretical findings.

2.1.1 *Models of Greenhouse Climate*

Most of the greenhouse climate models found in the literature are mechanistic physical models based on the mass and energy conservation laws. The models that arise from the application of these conservation laws take the form of mass and energy balances, mostly represented by systems of differential equations (or, alternatively, discrete-time difference equations).

The energy fluxes in the greenhouse are defined by the heat and radiation transfer through the cover and soil (Bot, 1993). These energy fluxes are mainly driven by the solar radiation and have temperatures as state variables (since the heat fluxes are temperature-driven). Although a thorough model of energy fluxes would need to know the temperature of the air and all the greenhouse's boundaries, that is, the covering and soil layers, (Takakura, 1993; Seginer and Kantz, 1986; López-Cruz

et al., 2008), many models limit themselves to measurements of the air temperature (e.g. Pasgianos et al. (2003); Ferreira et al. (2002); López-Cruz et al. (2007)).

In addition to the energy fluxes, two mass fluxes are of interest in the greenhouse climate: The water vapour and the CO_2 in the air. The importance of the water vapour balance is twofold: The water content in the air is the main driver of the plant transpiration (Stanghellini, 1988), and it plays an important role in the energy balance itself due to the phase change taking place during the evapotranspiration (Takakura, 1993; Seginer, 2002).

The greenhouse climate can thus be defined by the following 4 parameters (Fitz-Rodríguez et al., 2010; Bot, 1989b):

- Air temperature
- Relative humidity
- Solar radiation
- CO_2 concentration

As was mentioned before, some authors also include two aspects of the the soil in the models: The water content in the soil and the heat flux through it. Additionally, the air movement plays a role in the mass and energy transport, and can reach velocities leading to turbulent transport inside the greenhouse (Kittas and Bartzanas, 2007).

While the study of the CO_2 fluxes is useful to know (and ultimately ensure) the disponibility of this gas for the plant photosynthesis, there has been less research regarding this mass balance in greenhouses (at a canopy level). One reason is that the CO_2 control normally seeks to increase its concentration in the air (and not to lower it). In this sense, one main concern in CO_2 control is the price of the gas itself (Linker et al., 1998).

Avissar et al. (1985) describe the plant responses (stomatal conductance) to several physical factors, by means of the diagram reproduced in Figure 1. It can be seen that from the abovementioned climatic factors, only the response to (leaf) temperature has both upper and lower boundaries. In other words, both low and high temperatures decrease the stomatal conductance. That means that the radiation could theoretically be

increased arbitrarily without decrease of the stomatal conductance (which is closely related with the photosynthesis, as reviewed in Section 2.1.2).

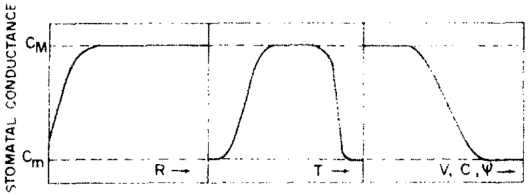


Figure 1: A schematic representation of the stomatal conductance response to global radiation (R), leaf temperature (T), vapour pressure difference between the leaf ant its ambient air (V), CO_2 ambient air concentration (C) and soil water potential (ψ). Reproduced from: Avissar et al. (1985)

Another interesting aspect of Figure 1 is that the depicted leaf temperature and vapour pressure difference are both closely associated with the air temperature and humidity. Indeed, the water content capacity of the air is a direct function of the temperature, which is thus involved in the calculation of relative humidity and vapour pressure difference.

From the point of view of greenhouse modelling, the most interesting variables are the air temperature and relative humidity, because: A) they are closely interrelated and influence each other directly, B) they can be easily measured, and C) they give information about the plant status. Hashimoto (1980) emphasized the *observability* and *controllability* of the plant leaf temperature and the relative humidity and set guidelines for plant oriented control of the climate[1].

In the last decades, mechanistic models of greenhouse temperature and relative humidity have been developed and calibrated for different greenhouse construction types and geographical locations (e.g. Bot and van Dixhoorn (1978); Udink ten Cate (1985); Tantau (1989); Seginer and Kantz (1986); Stanghellini (1987)). In a meta study, Boulard (2012) gives an overview of the publications dealing with greenhouse climate and plant modelling.

1 The observability and controllability of a system and its associated states are concepts from systems control theory, and can be reviewed for instance in Kuo (1995) and Ogata (2002).

Since it is well known that the air temperature and relative humidity are closely related, it has been proposed to model both fluxes (heat and water) simultaneously (Stanghellini, 1988), the connection point being the latent heat of vaporisation of the water (Boulard et al., 2002; Kittas and Bartzanas, 2007). This approach can be regarded as a modification of the *FAO - Penman-Monteith* equation (Steduto et al., 2012; Monteith and Unsworth, 2013), which has been widely used for irrigation scheduling in open-field agriculture (Allen et al., 1998). Work by several authors, like Stanghellini (1987), Seginer (2002) and Takakura et al. (2009) aims at a modified model for use under greenhouse covering. Another approach to the integration of temperature and humidity models makes use of the enthalpy as a state variable (Pasgianos et al., 2003; Schmidt, 2005; Salazar et al., 2010a; Bronchart et al., 2013).

One problem faced by mechanistic models is the difficulty to conceptually demarcate the system (the greenhouse) from its sorrounding environment, that is, to set the boundaries of the model.

To define the limits of the greenhouse system means to take the weather conditions into account. The weather has a strong influence on the greenhouse climate, to the point that some authors consider it to be the driving force behind it (Impron et al., 2007; Jones et al., 2016). However, under the greenhouse modelling framework, the weather is also commonly seen as a disturbance, an element that interferes with the realisation of a desired control strategy (Seginer, 1997; Pasgianos et al., 2003; Bennis et al., 2008; Van Straten and Van Henten, 2010). Another way of looking at the weather is to consider it as a group of variables interrelated with the greenhouse climate (Linker and Seginer, 2004), thus being part of the system and defining its boundary conditions (Bot, 1989b). One way to include (at least partially) these boundary conditions on mechanistic models is the use of theoretical values of the solar radiation, calculated for each time of the year and day in a particular geographical location (Thornley and France, 2007; Jones, 2014). The energy balance can thus take into account the flux of solar energy, even though all other fluxes depend on unknown temperatures and vapor pressures (Kimball, 1973).

The ventilation is a process that strongly underlines the greenhouse's boundaries: The air exchange rate between the greenhouse and its environment has implications on the heat as well as the water and CO_2 balances (Van Henten, 2003; Giacomelli et al., 2007). The calculation of

these fluxes is difficult even if the greenhouse air is assumed to be perfectly homogeneous (Roy et al., 2002). The loss of water vapour through ventilation reduces the relative humidity in the air, but also has a direct effect on the heat balance because the water carries latent heat to the outside of the greenhouse (Fuchs et al., 1997; Wang et al., 1999), particularly when forced ventilation is used (Katsoulas et al., 2002, 2007). Also, in facilities equipped with CO_2 fertilisation, this gas can get lost to the outside, where the concentration is normally lower than inside the greenhouse (Nederhoff, 1995; Kläring et al., 2007). The complex problems that arise from the air exchange through ventilation have been tackled using numerical techinques to solve the corresponding systems of differential equations. (*CFD: computational fluid dynamics*) (e.g. (Boulard et al., 2002; Kittas and Bartzanas, 2007)).

The so-called blackbox (descriptive) models can cope with some of the aforementioned difficulties in the mechanistic approach, because they rely on measurements and observations as a starting point to build up mathematical representations. This emphasis on measurements has proven useful in cases with limited instrumentation available, because these models automatically include all variables involved in the process being modelled. This model category comprises different approaches that relate variables, including regression and polinomial fitting (Blasco et al., 2007), autoregressive models like ARMA, ARIMA and ARX (Uchida-Frausto and Pieters, 2004; López-Cruz et al., 2007), ANN (Boulard et al., 1997; Fourati, 2014) and support vector machines (SVM) (Ljung, 2010). One drawback of these methods is their restricted extrapolation capabilities, which often bounds themodel to the system from which the building data came (Seginer, 1997; Linker and Seginer, 2004).

The autorregressive models reported in the literature (Uchida-Frausto et al., 2003; Uchida-Frausto and Pieters, 2004; López-Cruz et al., 2007; Patil et al., 2008) treat the air temperature and relative humidity as self-dependent time-series, rather than a result of the external climate. In these cases, all other variables are considered *exogenous*, meaning that they are external, additional to the variable of interest that is calculated using autorregression.

Descriptive models of the greenhouse climate, due to their empirical nature, implicitly include the effects of the outside weather, but also the

role of the plants growing inside: It is a well-known fact that the plants actively modify the environment they live in.

Monteith and Unsworth (2013) express this idea as follows:

> "The presence of an organism modifies the environment to which it is exposed, so that the physical stimulus received *from* the environment is partly determined by the physiological response *to* the environment."

Similarly, von Zabeltitz (2011) point at an integrated technical system when it comes to the greenhouse technical construction and equipment:

> "A greenhouse structure with a light transmittance (through framework and cladding material), ventilation, heating, cooling and protection from pest insects by screening, as well as all influences of management, irrigation, fertilisation, water quality, physical and biological plant protection, has to be considered as an integrated system."

2.1.2 *Models of Plant Transpiration and Photosynthesis*

The plants modify the environment in which they live as part of an active response to its stimuli. Although these modifications take place in the root as well as in the air environment, the latter are more important for greenhouse horticulture, since the main task of a greenhouse covering is to demarcate a confined air volume for the plants to grow. The confined air can help to protect the plants from plagues and birds, but also make it possible to use technical equipment to modify the physical properties of the air, mainly temperature, humidity and CO_2 concentration, as described in Section 2.1.1.

One clear example of plants modifying their environment is given by the transpiration. The transpiration is closely related to the amount of water present in the air, i.e. how wet or how dry the sorrounding air is. In cases where the air is dry, the plants transpire more, which in turn leads to an increase of air moisture. The whole process is affected by the availability of water in the soil, the transpiration itself being driven by the water potential in the diferent parts of the soil-plant-air continuum,

or, more precisely, the difference of water potential between the soil (or plant substrate) and the air (Kramer and Boyer, 1995; Jones, 2014).

The determination of the plant transpiration (be it by means of calculations or measurements) is important to define irrigation strategies that ensure an adequate water supply (Jones and Tardieu, 1998). Allen et al. (1998) describe the rationale behind irrigation scheduling as a mass balance of water, where the *evapotranspiration* (defined as the sum of direct evaporation from the soil and transpiration by the plants) represents a loss term that needs to be restored via irrigation. The method described by the authors is also the standard practice recommended by the FAO (*Food and Agriculture Organization of the United Nations*) to determine the water needs of open-field crops, and is based on the reference evapotranspiration (ET_0) equation by Penman and Monteith (Monteith and Unsworth, 2013). This Penman-Monteith equation ($P-M$) takes into account the latent heat flux involved in the phase change that occurs during evaporation, and was derived from the energy balance on an open water surface, adapted to a big crop surface later on[2].

The aforementioned energy balance is divided in two parts: a radiation and an aerodynamic terms, being the latter directly related with the vapour pressure difference between the leaves and the surrounding air. The $P-M$ equation is shown in its simplified form in Equation (2.1.2.1) (after Kittas et al. (1999)), where TR is the transpiration rate, R_s the solar radiation on the crop and VPD the vapour pressure deficit.

$$TR = A \cdot R_s + B \cdot VPD \tag{2.1.2.1}$$

Several differences between open-air and greenhouse cultivation must be taken into account in order to use the $P-M$ equation to calculate the evapotranspiration of greenhouse crops. Among them are the fact that in most greenhouses the direct evaporation from the soil is very low (as it is commonly sealed out with plastic or concrete) and that the canopy structure is often different from the extensive, dense surfaces of homogeneous height for which the model was originally developed.

2 The reference crop is defined as *an extensive surface of green grass of uniform height, actively growing and adequately watered* (Allen et al., 1998)

Thornley and France (2007) explain the role of the canopy structure using the *Big Leaf* concept:

> [In the classical $P-M$ equation,] *"the canopy is treated as a single or big leaf. Temperatures at sites in the canopy where water evaporates (sub-stomatal cavity) and where canopy-to-air heat transfer takes place (leaf surface) are assumed the same. Heat storage by the canopy and heat-producing metabolism in the canopy are ignored."*

Seginer (2002) agrees and thus recommends the use of the leaf temperature as starting point to the calculation of greenhouse transpiration, since it is in the stomata where the water phase change takes place.

In adapting the $P-M$ equations to greenhouse conditions, many authors (e.g. (Stanghellini, 1987; Jolliet and Bailey, 1992; Kittas et al., 1999)) have incorporated the Leaf Area Index in the calculation of the A and B coefficients in Equation (2.1.2.1). The Leaf Area Index (LAI) is a dimensionless parameter relating the area of leaves to that of the underlying ground surface, i.e. it is a measure of soil covering (Allen et al., 1998). The exponential relation between the A (radiation coefficient) and the LAI follows the results on light interception (vertical exponential decrease) inside a canopy originally reported in the seminal work by Monsi and Saeki (2005) (Originally published as: Monsi and Saeki (1953)).

An additional aspect regarding transpiration of greenhouse crops is the arguably low velocity of the air, which has led some authors to suggest the elimination of the second (aerodynamic) term in Equation (2.1.2.1), as the so-called *decoupled greenhouse* is assumed to be effectively isolated from the outside weather (Jones and Tardieu, 1998). Under such conditions, the air movement has little effect on transpiration (because the thickness of the boundary layer next to the leaves depends on the wind velocity), the mass transport processes for both H_2O and CO_2 can be assumed to be driven primarily by diffusion (Stanghellini, 1988), and thus the main drive for transpiration is the solar radiation (Stanghellini, 1987). Seginer (2002) regards this assumption as *too severe* because, even though it can hold for tightly closed greenhouses (e.g. closed or semiclosed greenhouses in temperate climates), it does not apply to higher ventilation rates or when forced ventilation is used. Baille et al. (1994) also found significant differences in the radiation and aerodynamic terms, when comparing results between crops and control actions.

Takakura et al. (2009) have also criticized the sole use of radiation for irrigation control, which they identify as an *european practice of irrigation*, and recommend the use of direct measurements, in addition with an energy-balance equation (first published by Takakura et al. (2005)), to directly calculate the evapotranspiration of sparse canopies. The sensor proposed by the authors (further developed by Akutsu et al. (2015); Takakura et al. (2017)) aims at measuring most energy fluxes by a combination of radiation sensors, while stressing the importance of the surface temperature of the canopy as a whole. After comparing the models by Stanghellini (1987) and Takakura et al. (2005) with the $P - M$ equation, Villarreal-Guerrero et al. (2012a) concluded that the first can be slightly more accurate if the LAI as well as internal resistances can be appropiately incorporated, while the second is more robust in general, but is sensible to the sensor placement and position.

The importance of direct measurements of plant transpiration, stressed by authors like Takakura et al. (2005), is also of importance to calibrate and evaluate the estimations made by all matemathical models.

The main method for measuring plant transpiration is the lysimeter, used to make a water mass balance by measuring water input and output from the rhizosphere (Allen et al., 1998; Jones, 2014). The weighing lysimeters used in open field are costly devices that take the water storage in the soil and plants into account by constantly measuring the weight of a block of planted soil (Kramer and Boyer, 1995). For horticultural crops grown in greenhouses, small weighing scales have been used to monitor single plants, or small groups of them (e.g. Stanghellini and de Jong (1995); Sánchez et al. (2012)). This method is regarded as one of the most accurate for measuring of transpiration. However, the installation of weighing devices makes it difficult for use outside of research facilities. Instead, a simple water balance can be made in greenhouses by making the assumption that the water storage in the plants (for short periods) and in the soil (because of a small volume) is negligible.

A different, less common approach, aims at measuring (or calculating) the water flow through the plant. Two examples of this approach are measurements based on stem diameter variations (Steppe et al., 2008; Fernández and Cuevas, 2010) and water output from the leaves (Ton et al., 2004; Schmidt, 1992).

As the assessment of plant transpiration involves the measurement of water (either liquid or vapour), measurements of CO_2 are needed to determine plant photosynthesis. In general, these measurements are technically more difficult to carry than those of water. Jones (2014) classifies the methods for measuring CO_2 exchange as follows:

- Growth analysis: use of long-term growth measures, most commonly destructive dry-mass weight measurement, to infer back the assimilation over a large period
- Use of isotopic tracers: use marked carbon atoms to estimate the fluxes in small scale
- Net gas exchange: measurements of CO_2 or O_2 to estimate photosynthesis and respiration by means of mass balances

Many devices and methods for making photosynthesis measurements have been historically used to investigate the chemical and physiological pathways of this process, and thus involve labour-intensive and destructive methods carried out in the laboratory (Hunt, 2003; Millan-Almaraz et al., 2009). These devices and methods fall therefore more into the field of botanical than agricultural science.

From the methods mentioned before, the two most prevalent in agriculture and horticulture are those based on a long-term assimilation and those based on gas exchange measurements, for short-term measurements on living plants. In agricultural terms, the former can be seen as a means to relate the photosynthesis to the total crop growth (and yield), while the latter would be useful for the evaluation of short-term management and growing techniques over time.

The models of photosynthesis can be also be classified like the measurements: a first kind of mathematical models have been mainly oriented towards the botanical side of this plant process, as a means to investigate its physiology and biochemistry. The most reputable among these models is the one developed by Farquhar et al. (1980), which has become the standard in the topic. It links the gas exchange kinetics (O_2 and CO_2) with the leaf biochemistry (Farquhar et al., 2001) by splitting the photosynthetic response curve into three separate segments, defined by three corresponding biochemical processes (RuBisCO-limitation, RuBP-limitation and TPU-limitation) (Jones, 2014).

A second group of models focuses on plant growth and dry-matter partition (e.g. Heuvelink (1995); Marcelis et al. (1998)). This kind of models strongly rely on measurements (rather than theoretical first principles) to analise the plant development over longer periods of time (weeks to months). Perhaps the best-known model of this kind, developed for tomatoes, is TOMGRO, by Dayan et al. (1993). While TOMGRO is not a pure photosynthesis model, it contemplates the calculation of the instantaneous assimilation rate (of CO_2) as a step to the computation (by integration) of the daily gross assimilation.

One difference between transpiration and photosynthesis is the way in which the gas exchange takes place: the transpiration defines directly a flow of H_2O in a single direction, namely out of the stomata opening. On the other hand, the flow of CO_2 (as well as O_2) is affected by the photosynthesis, respiration and photorrespiration (Hunt, 2003). Since both CO_2 and O_2 are produced and absorbed simultaneously, the net photosynthesis rate represents the balance between CO_2 uptake and O_2 release in photosynthesis, against CO_2 release and O_2 uptake in both respiration and photorrespiration (Hunt, 2003). The straigthtforward relationship of the stomata opening with the vapour pressure deficit (Franks et al., 1997) also links both plant processes with the physical conditions of the sorrounding air.

The relation of the photosynthesis with the changing conditions of the air opens the possibility to influence this plant process through climate control in greenhouses. This possibility has motivated the development of measuring devices for continuous use on living plants. Following Udink ten Cate et al. (1978), the incorporation of continuous plant measurements (phytometric signals) to the greenhouse controller has been called *the speaking plant approach*.

Most continuous-operation systems reported in the literature measure climatic variables and use them in conjunction with mechanistic models to calculate the transpiration and photosynthesis (Ehret et al., 2001). Long-term measurements of photosynthesis and transpiration are mostly carried out using growing chambers enclosing full plants (e.g. Maclean et al. (2012)) or make labour-intensive measurements per hand (e.g. Kläring and Krumbein (2013)). Methods that mechanically attach the measurement device to the leaves are better suited for autmated climate

control, although they are expensive and need to be relocated as the plants grow (Schmidt, 1992; Schmidt et al., 2014).

Summarising, one can identify two approaches to the determination of transpiration and photosynthesis. A first approach, commonly used in agriculture and horticulture involves single-pointed (not seldom destructive) measurements of plant and soil samples, and strongly relies on weather variables and bulky devices (as lysimeters, for example). The models developed make assumptions of homogeneity (*big leaf models*) and work with mass and energy balances that aim at predicting plant growth and yield over periods ranging from days to months. By contrast, a second approach uses mechanistic models of the plant physiology, supported by measurements taken with costly, laboratory-like instrumentation. This approach carefully inspects the plant processes in single leaves (or parts of them) in a time scale of minutes. The literature shows a trend in horticultural science towards a more detailed monitoring of whole plant canopies as support for fast automation systems.

2.2 ARTIFICIAL NEURAL NETWORKS IN GREENHOUSES

Artificial Intelligence (AI) as a discipline deals with the main objective of building machines that show intelligent behaviour (Nilsson, 2009). Even though it is a comparatively young science[3], it comprises a vast field of study and has an active scientific community. Therefore, even a modest introduction to the topic would fall beyond the scope of the present work. Nonetheless, the following lines are included for the sake of giving a frame of reference to Section 2.2.1 and Section 2.2.2.

The sole process of building intelligent machines brings about debates and discoveries that enrich our understanding of intelligence itself. As a consequence of this broad aim, the scope and methods of AI span from the engineering to philosophy, linguistics and psychology (Green, 1996; Bly and Rumelhart, 1999). Pfeifer and Iida (2004) put this as follows:

> *"While originally artificial intelligence was clearly a computational discipline, dominated by computer science, cognitive*

3 It was officially founded in 1956, when the name and main objectives were presented in two conferences in the USA (Pfeifer and Scheier, 2001).

> *psychology, linguistics and philosophy, it has turned into a multidisciplinary field requiring the cooperation and talents of many other fields such as biology, neuroscience, engineering (electronic and mechanical), robotics, biomechanics, material sciences, and dynamical systems."*

In their search for intelligence, AI-researchers have created a variety of techniques and methods that consider different descriptions of intelligence and the corresponding ways to build it. A non-extensive list of techiques and methods that emanated from or were inspired by AI-research includes:

- fuzzy logic
- knowledge-based systems (also: expert systems)
- artificial neural networks
- swarm intelligence
- genetic algorithms
- embodied robotics

Independently from the role that the mentioned techniques have played in the search for intelligence, many of them have found appplications in other areas of science and engineering. This is also the case with artificial neural networks, as will be further discussed in this chapter.

As a closing note, it can be mentioned that in recent years some AI-researchers proposed a shift of the basis for mental behavior from abstract symbols toward the situated body, giving more importance to physical interaction in the form of perception and action (Pecher and Zwaan, 2005; Pfeifer and Iida, 2004; Pfeifer and Bongard, 2006). It can be expected that this new approach contributes with new techniques and methodologies that can both deepen our understanding of intelligence, but also strengthen and enrich other areas of science and engineering.

2.2.1 *Artificial Neural Networks*

Artificial neural networks (ANN) are mathematical constructions (algorithms, programs, sets of equations) that were first developed imitating the way in which neurons in the nervous system interact with each other. This means that their operation model was originally inspired by knowledge about how biological neurons work (Ripley, 2006). Most common neural cells found in animals have the structure shown in Figure 2, where two main parts are visible: The neuron's core features on the one side a number of relatively short extentions called *dendrites* and, on the other, a single long extention, the *axon*. Electrical stimuli from several sources arrive at a single cell via the dendrites, and the cell itself is able to produce an electrical discharge (*fires*) through the axon (Horner and Kühn, 1998). The occurrence of the latter depends on the incoming stimuli and on physiological parameters pertaining each particular cell (Nilsson, 2009).

An man-made element that resembles the functioning of a single biological neuron is the artificial neuron. While there are different types of artificial neurons, the most broadly known is the *perceptron*.

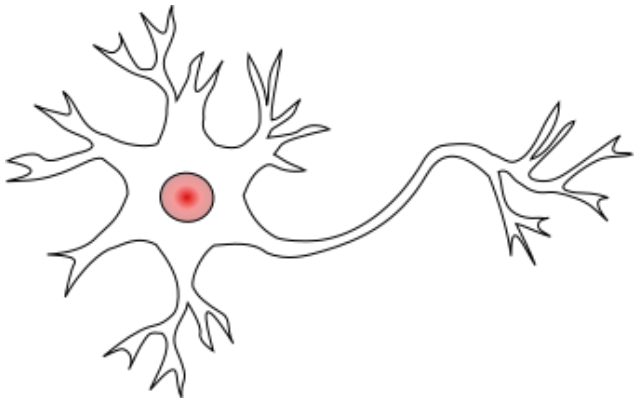


Figure 2: Diagram of a neural cell. On the left is the nucleus with the dendrites. The axon extends to the right. Modified from: Jonathan Haas. https://commons.wikimedia.org/wiki/File:Pseudounipolar_bipolar_neurons.svg. Consulted on 30-Jun-2017.

The perceptron was developed in 1957 by Frank Rosenblatt, and was originally conceived as a hardware device, with physical connections to conduct electricity. A perceptron constitutes a single calculation unit, with a variable number of inputs and a single output, that mimick the biological neuron: the information flows from the inputs into the nucleus, where a calculation is made and eventually an output impulse is triggered. The perceptron was further developed as a piece of software, more concretely, as a calculation algorithm (Nilsson, 2009).

The calculations carried out by a single perceptron unit are relatively simple, and can be divided into 2 parts (Russell and Norvig, 2010):

1. determination of a weighted sum of the inputs,

   and

2. non-linear transformation of the obtained result

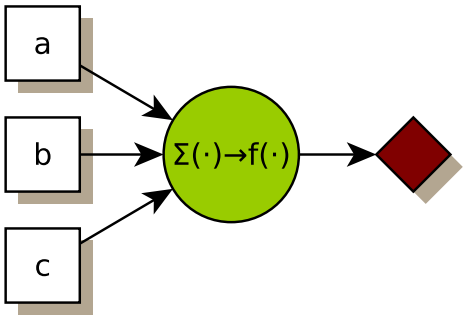


Figure 3: A single perceptron unit, featuring three inputs and a single output. The weights associated with the inputs are not shown. The activation function f is applied to the weighted sum of the inputs.

The first part represents a *linear combination of the inputs*, a basic operation in linear algebra. The second is of capital importance for the approximation capabilities of a perceptron, because it allows to estimate very complicated functions, as will be described below in this section. The non-linear function applied by the second part of a perceptron is called the *activation function*, and is inspired by the firing process taking place in biological neurons (Horner and Kühn, 1998). A most common family

of mathematical functions that have been used for this purpose are the *s*-shaped *sigmoidal functions*. These functions saturate at both high and low values of the function domain (although at different values for each), and are nearly linear in the center part. Two examples of sigmoid functions commonly used as activation functions in neural networks are shown in Figure 4. One characteristic of these functions is that their derivative can be expressed in terms of the original function, which is computationally efficient and thus desirable for the learning algorithm described below. For the activation function $f(x) = tanh(x)$, for instance:

$$\frac{d\left(f(x)\right)}{dx} = 1 - tanh^2(x) = 1 - \left(f(x)\right)^2$$

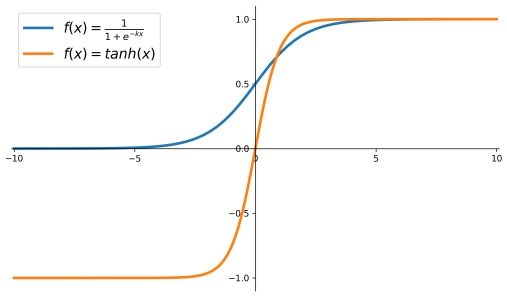


Figure 4: Two sigmoid functions commonly used as activation functions for neural networks.

The possibility to use individual units, perceptrons, as building blocks to make a complex network, together with the non-linear activation function of each node, made it possible to use such an ensemble for complicated calculations involving non-linear systems. In such cases, it is said that a complex behaviour *emerges* from the combination of a number of relatively simple elements. Because of this emphasis on the connections themselves, this approach to problem solving has been called *connectionism* within the field of AI (Russell and Norvig, 2010; Dawson, 2008).

The connection of a node's output to another node's input gives rise to a layered structure, particularly in cases where a group of units

at one step of the calculation is connected to a subsequent group. This architecture is called a *multi-layer perceptron* network (MLP), and represents the most common form of ANN. In its canonical form, where all nodes in one layer are connected to all nodes in the next one, it is said to be *full-connected*. 3 fully-connected multilayer perceptron networks can be seen in Figure 14 on page 58, as examples of the models developed in this research.

The multilayer perceptron networks have been known for their approximation capabilities, and are said to be *universal approximators*, which means that it is theoretically possible to approximate any arbitrary function with a network of enough complexity (number of nodes and layers). This is known as the *universal approximation theorem* (Cybenko, 1989; Hornik et al., 1989). Russell and Norvig (2010) cite the work by Cybenko (1989) in saying that *two hidden layers are enough to represent any function and a single layer is enough to represent any continuous function*. This agrees with the design recommendations given by Masters (1993) of using a single hidden layer, and only a second one for very difficult problems, if the single-layered network does not work properly.

In order to approximate a function (often represented by a set of data, rather than a symbolic expression), a neural network needs to be calibrated, a process called *training* or *learning*.

The learning process of a perceptron implies finding fitting values of the weights affecting the inputs (Riedmiller, 1994), that is to say, the callibration of the weighted sum of the inputs to the unit. The perceptron is thusly *trained*. While the training can be done by using any optimization technique, the most widespread training algorithm is the *backpropagation of error* (e.g. Werbos (1988); Bishop (1995); Nilsson (2009)).

The backpropagation algorithm performs a *supervised* training of the network, because it iterates over a dataset consisting of a list of inputs paired with the corresponding expected output. These expected output data must be known beforehand (Bishop, 1995). For each input-output-pair, the algorithm calculates the difference between the network's output, calculated with the current values of the weights, and the expected output. The weights are subsequently adjusted to minimize this difference, according to its derivative, a process also called *gradient descent* (Menzel, 1998; Riedmiller, 1994). This is the reason why functions

whose derivative is easy to calculate represent an attractive choice as activation functions (Nilsson, 2009). The algorithm's name underscores two directions of information flow: during standard run time, the information flows from the input to the output paths of the perceptron, a process in which the output value is also calculated. This is called a *feedforward run*. During backpropagation of the error, that is, during the training process, the information on the error i.e., the difference between calculation and expected result, flows back from the outputs towards the inputs, and is used to update the weights according to it. If an ANN is composed of several individual perceptrons (MLP), as is the case for all practical applications, the algorithm is sequentially applied to each of them. The adjustment of the weights is repeated once the complete set of training data is used. Each training iteration receives the name of *trainig epoch*.

Because the backpropagation algorithm is an optimisation algorithm that tries to minimise the difference between the network's expected and actual output, it is prone to *overfit* the data. This means, that it could find a set of weights that effectively minimises the error, but performs poorly when presented with new data records, not seen during the training process. This phenomenon is a consequence of the network's complexity, and should not happen if its dimensions (number of nodes and layers) are selected in conformity with the task's own complexity (Masters, 1993; Ripley, 2006; Goodfellow et al., 2016). In general terms, a simple problem should need a less complex network (i.e. less nodes), and vice versa. In fact, function overfit can occur with all function-approximation techniques, not only neural networks. As an example, when a small set of data is used to calibrate the coefficients of a high-order polynomial, it is possible that the resulting values reflect the variability of the data with too much detail, not being able to capture the underlying processes that brought about the data (Bishop, 1995; Goodfellow et al., 2016).

The backpropagation algorithm requires that the networks' weights are different from 0. For this reason, it is needed to set their initial values to some fit number to start the optimisation process. It is a common practice to randomly set their initial values to small numbers. Regarding this randomness, Bishop (1995) points that:

> "Since a particular training run is sensitive to the initial conditions for the weights, it is common practice to train a partic-

> ular network many times using different weight initializations. This leads to a set of different networks whose generalization performance can be compared by making use of independent data. In this case it is possible to keep the best network and simply discard the remainder."

A consequence of the random initialisation of the weights is therefore that the training is non-deterministic, and each training process can yield different results as the weights' values converge towards different optimal values. Ehret et al. (2011) discusses how this random initialisation gives rise to a group of models with similar performance, rather than a single *best* neural network. In order to account for this randomness when comparing and selecting networks' architectures, the authors repeated the training process 30 times.

Since there are no general rules to find the best network's complexity (architecture) for a given problem (Masters, 1993; Ripley, 2006; Goodfellow et al., 2016), the network's architecture design implies a trial-and-error search. This search is stochastic and can be influenced by the weights' initialisation described above.

A matemathical model's *parameters* are those values that need to be calibrated for the model to perform a desired task. In a polynomial model, for example, they are the coefficients of the individual terms. In a neural network, the parameters are the network's weights, which are set by the backpropagation algorithm (or any other optimisation technique). Conversely, a model's *metaparameters* (also called *hyperparameters*, Goodfellow et al. (2016)) are those values related with the matemathical form of the model, it's architecture and complexity. In ANN, the metaparameters include the number of nodes and layers, as well as the activation function.

Since the calibration of a network's parameters (training) and the selection of its metaparameters (selection of the architecture) represent two independent processes, it is recommended to use different groups of data to carry them on.

The most widespread terminology (e.g. Bishop (1995); Ripley (2006); Goodfellow et al. (2016)) for the datasets needed in neural modelling is then presented by Ripley (2006) as follows:

- Training set: A set of examples used for learning, that is to fit the parameters [i.e. weights] of the classifier.

- Validation set: A set of examples used to tune the meta-parameters [i.e., architecture, not weights] of a classifier, for example to choose the number of hidden units in a neural network.

- Test set: A set of examples used only to assess the performance [generalization] of a fully-specified classifier.

It can be noticed that the definitions above refer to the neural networks as *classifiers*. There are two big applications of neural networks: classification and regression (Russell and Norvig, 2010), being the latter the one that refers to function approximation as discussed previously in this section. When neural networks are used for classification, they are presented with input data that belongs to a category, which represents the network's output. After training, it is the network's task to trigger the corresponding output node, and ideally give an unanbiguous class to which each set of inputs belongs.

In addition to the common multilayer perceptron, a number of neural networks are reported in the literature. Mostly, they differ in the connections between individual nodes and the activation functions used. Among them are the radial basis function, which uses a spherically symetric function (such as a Gaussian density function) instead of a sigmoid as activation function (Cheng and Titterington, 1994; Linker and Seginer, 2004).

Other type of network is the recurrent neural network, which features output connections that also serve as inputs to the same nodes (Oussar and Dreyfus, 2001; Schmidhuber, 2015). A wide variety of topologies arise from differences in the way that the recurrent connections are made: a recurrent connection can be fed back to the same node, to all the nodes in the same layer, or to the layers located before (e.g. Menezes and Barreto (2008)).

2.2.2 *Applications in Greenhouse Horticulture*

There has been a long-lasting interest in the use of neural networks for applications in greenhouse horticulture. The motivation originates from the complexity of the greenhouse biosystem, which makes it difficult to model with mechanistic techniques (e.g. Challa and van Straten (1993); Hashimoto (1993); Morimoto and Hashimoto (1996); Fourati and Chtourou (2007)).

Most published work regarding ANN and greenhouses is devoted to the simulation of the internal climate as a function of the outside weather conditions.

Seginer and Sher (1993) proposed to use already trained neural networks to search for an optimum daily set point temperature. For this purpose, they used several years of daily average data to train the networks, and tested them subsequently using other year's data. In addition to the average outside temperature and solar radiation, the authors included measures of crop growth as model inputs: number of days since transplanting, number of plant nodes, dry weight and fruit fresh weight. In a similar approach, Linker et al. (1998) trained two networks to be used as a model of air temperature and CO_2 concentration as a function of the solar radiation and temperature outside, as well as the state of two greenhouse actuators (ventilators and thermal screen). The trained networks were used to perform an exhaustive search for optimum set points for both variables. In this case, the calculations were done in a time scale of minutes, using two months of data measurements.

In the previous examples, the trained networks were first used to model the greenhouse system as a whole, and then to search for set points for an already existing control system. Therefore, their intended use falls into the category of *optimal control*, as described by Udink ten Cate and Challa (1984).

When the weather conditions outside the greenhouse are used as inputs to calculate the climate inside it, they are sometimes called *exogenous variables*, to indicate their influence over the variable being modelled as an autorregressive process. Therefore the acronym NNARX used with neural networks means that the inputs include the same variable being modelled, as well as other, not present in the outputs. The NNARX

models represent the neural counterpart of the ARX mentioned in Section 2.1.1. This model architecture has been reported to give good estimations of the current air temperature inside a greenhouse, using previuos values of the variable and the weather outside (Uchida-Frausto and Pieters, 2004; Patil et al., 2008). Similar networks were used by Fitz-Rodríguez et al. (2012) to predict air temperature and relative humidity in 10-seconds time steps, as a function of the external climate and the actuator operation (ventilation opening and fogging).

The forecast of climate values in the near-future horizont has also been addressed by He et al. (2007), who used the external climate to estimate the temperature and relative humidity inside a greenhouse, one time step (30 minutes) in advance. Similarly, Salazar-Moreno et al. (2008) developed a single network to estimate the temperature and relative humidity inside a greenhouse, one step in advance (5 minutes). Both authors recommend to use $(2 \cdot n + 1)$ nodes in a single hidden layer of the neural network, where n represents the number of inputs. Another parallel between both research works is the length of the dataset used: 15 days of measurements in the first case, and 8 days in the second. By contrast, Dariouchy et al. (2009) developed a neural network for one-step-ahead prediction (15 minutes) of greenhouse temperature and relative humidity by adding hidden nodes and testing repeatedly. For 5 inputs (outside climate: temperature, relative humidity, solar radiation, wind direction and wind velocity), they found the best architecture to have 6 hidden nodes.

As stated in Section 2.2.1, the relation between the network's complexity, the complexity of the problem to be tackled and the amount of data available is a sensitive step in model design.

The issue of the number of hidden nodes in neural networks (and in general, the selection of metaparameters) is discussed in the framework of greenhouses by Ehret et al. (2008) and Ehret et al. (2011). To assess the effect of the random initialisation of the network's weights, the authors repeated the model training 30 times for each architecture tested. In general, for regression tasks of yield, growth and water use (daily values), Ehret et al. (2011) found a number of nodes slightly lower than the number of inputs to perform better.

Compared with the greenhouse climate, less research work is reported about neural models of plant processes, being the abovementioned

yield prediction (Ehret et al., 2011) one example of them. The prediction of weekly yield has also being adressed by Lin and Hill (2008) on a sweet pepper crop and Salazar et al. (2015) on tomato. Due to the time scale, both works rely on relatively small datasets (156 and 26 records, respectively). By contrast, the neural networks (2 hidden layers with recurrent connections) for tomato yield prediction reported by Fitz-Rodríguez and A. Giacomelli (2009) were built with a substantially bigger database consisting of 2172 weekly records from two production cycles and four greenhouse sections.

Another research work featuring different cultivation periods (of a soilless tomato culture) and two different greenhouses is reported by Seginer et al. (1996). The authors used a dataset of 3,076 data records (half-hour averages) of weather values and greenhouse actuators (heating flux, sum of ventilator opening angles and fraction of time when the misting system was turned on) to predict the solar radiation, dry and wet bulb temperature inside the two greenhouses. The greenhouses were located in Avignon, France, and Salt Lake City, USA, making this the most comprehensive study found in the matter. The work builds upon that by Seginer et al. (1994), who compared neural models of climate trained with data from Avignon, France, and Silsoe, UK. The authors warn about using neural models outside their training domain, for example, using in winter a model trained in summer, because the domain of these models is defined by the training data used. Also, both articles suggest to use the relative values of the network's weights as a measure of input importance and to inform further experiments. In all cases, a single network was created for each greenhouse, using the data from a single location.

Fewer neural models are reported that deal with short term plant processes than with yield or climate. Zee and Bubenheim (1997) used a water balance mechanistic model to generate transpiration data, as expected inside a growing chamber. The data was subsequently used to train a neural network, with the transpiration being a function of the air and canopy temperatures, relative humidity and a *"resistance value identifying the type of plant"* (not specified). Shimizu et al. (2004) developed 6 neural models to calculate the shoot-tip temperature in 6 species of ornamental pot plants. As inputs, they used the dry and wet-bulb temperatures, as well as the greenhouse glazing temperature and the solar radiation. The shoot tip temperature was measured using a thermocouple for about a month (the actual periods being slightly

different for each plant), and the models were developed with a time step of 10 minutes. Using one week of measurements, Salazar-Moreno et al. (2011) created two neural networks for prediction of CO_2 concentration and photosynthesis rate. The photosynthesis values, measured with a gas exchange device every 15 minutes, were interpolated to fit the time step of 5 minutes that was used for the predictions (1 and 2 steps ahead). In an analysis of the input importance, the authors concluded that relative humidity was a very important input to the photosynthesis model, and explain *"that water vapor and CO_2 share the same pathway (stomata) in the plant, so an increase in plant water stress, the stomata closes and less CO_2 can be absorbed by the plant causing a reduction in photosynthesis"*. Additionally, the same authors used a month worth of data to build a neural network to predict the air temperature as a function of current and past values, as well as the current weather conditions outside the greenhouse.

3

MATERIALS AND METHODS

3.1 DATA COLLECTION

This section describes the greenhouses used in this research, along with their technical equipment. The data used for the simulation experiments is described in detail, too. These data came from three main sources:

- Direct measurements taken inside the two experimental greenhouses:
  - Climatic variables
  - Phytometric signals
- Direct measurements from an automated meteorological station
- Calculations and regression models derived from measurements:
  - Solar coordinates
  - Leaf Area Index

3.1.1 *Greenhouse description*

The implementation of the ANN models made use of data measurements originated in two greenhouses. These two greenhouses were built for the ZINEG Project[1], which ran from 2009 through 2014 in several institutions throughout Germany and aimed at reducing the energy consumption in these plant production systems. For the present research work, the data was taken at the Berlin facility, comprising 3 complete cultivation periods

1 In book: Niedrigenergiegewächshäuser. Ergebnisse des ZINEG-Verbundprojektes, Edition: 509, Chapter: 2, Publisher: KTBL, Darmstadt, Editors: KTBL. (In German); http://www.zineg.net/ZINEG_E/.

from 2011 to 2013. This facility consists on two similar Venlo-type greenhouses, shown in Figure 5. A summary of the greenhouse construction data is shown in Table 1.

Figure 5: ZINEG Greenhouses in Berlin. Left: Collector greenhouse. Right: Reference greenhouse

While both greehouses share location, size and construction style, they differe mainly in that the *Collector Greenhouse* was equipped with a better isolation (double glass walls, extra thermal screen under the roof) and a passive cooling system based on heat exchangers installed under the roof (Figure 6). Thanks to the enhanced isolation and the cooling system, the ventilation openings in the collector greenhouse could be kept closed for longer periods than those in the *Reference Greenhouse*. This is called a *Closed* or *Semi-closed* operation mode. As a side effect of the cooling system and the closed operation mode, the collector greenhouse was able to yield thermal energy to a storage tank installed nearby. The stored thermal energy could be used to either cool down or heat up the greenhouses by using a reversible heat pump.

Characteristic	Value
Location	Berlin, Germany $(52°28'02''N, 13°17'57''E)$
Construction type	Venlo-type Glasshouses
Ground area	$307m^2$
Sidewall height	$6m$
Ridge height	$6.7m$
Isolation	Double glass on the walls, single glass on the roof. Horizontal thermal screen under the roof in both greenhouses. One additional thermal screen in the collector greenhouse (at the side walls and under the roof).
Culture	*Solanum lycopersicum* in elevated rockwool slabs.
Irrigation	Closed hydroponic system with drip irrigation.
Heating	Hot water-pipe rail heating system on the floor. Tubular-film blowers under the benches. Aluminum heat exchangers in the canopy.
Cooling	Fin-tube heat exchanger under the roof for condensation cooling (only in the colector greenhouse).

Table 1: Physical and technical description of the greenhouses.

(a) (b)

Figure 6: Two technical systems to reduce energy consumption in the collector greenhouse. (a) Fin-tubes used as heat exchangers for cooling purposes. (b) Double horizontal screening under the roof.

Aside from saving energy (both by reducing the heating needs and by yielding thermal energy from the cooling system), the semi-closed operation of the collector greenhouse extended the periods where the CO_2 enrichment system could be used (the CO_2 enrichment was constrained to the periods when the windows were closed).

Both greenhouses were used for growing tomato plants (*Solanum lycopersicum* L.), as shown in Table 2. During year 2012, half of each greenhouse's cultivable surface was used to grow a different variety. In each greenhouse, 480 plants were soilless grown in rockwool slabs using a closed hydroponic system. The plants were organised in 5 double lines, plus a single line in each outer side. Every line had 40 plants, at 0.5m intervals.

A central computer system recorded data from all sensors and actuators both from the experimental greenhouses as well as from the weather station installed outside. This system was called *Plantputer System*[2] (Steinbeis GmbH & Co. KG for Technology Transfer; Berlin, Germany), and was responsible for the data collection tasks, as well as the automatic control of the actuators in the greenhouses. The Plantputer System included software to set up control strategies for climate and irrigation (automatic as well manual), and also managed the storage of all control variables and sensor measurements. A series of electronic printed ciruit boards was installed in each greenhouse, both for data acquisition and for activation of the actuator outputs.

The Plantputer software stored all measurements in a series of raw-text data files, in CSV format. The data from different sources was correspondingly stored in different files. From the broad range of stored data, only the meteorological measurements from open air, the climatic

2 http://www.plantputer.de/

Year	Variety
2011	Pannovy
2012	Komett, Encore
2013	Pannovy

Table 2: Tomato varieties grown during the research period.

measurements from inside the greenhouse and the phytometric measurements taken directly from the plants were of interest for this research. These measurements will be described in Sections 3.1.2 and 3.1.3.

3.1.2 *Climate and Weather*

This work used three climatic variables from inside the greenhouses: Air temperature [$^{\circ}C$], relative humidity [%] and CO_2 concentration [ppm]. There were 6 points inside each greenhouse where the air temperature and relative humidity were measured. Five of these points were located at 2m height and the sixth at 6m height. The Plantputer System stored all single sensor's measurements (which were taken every 30 seconds, approximately), but also calculated and stored their mean value, which was the value used for the calculations in this research (mean value every 5 minutes). In other words, the *perfectly stirred tank assumption* (Roy et al., 2002) held (no spatial gradients of temperature, relative humidity and CO_2 concentration were taken into account). The resistive temperature sensors (PT1000), the capacitive humidity sensors as well as the infrared-absorption CO_2 probes were installed in white, ventilated (continous air flux in suction) metal cases to improve accuracy by avoiding air stagnation and protection from direct radiation (Figure 8).

The Plantputer System also granted access to meteorological data, measured with an automatic station located at approximately 80 meters from the greenhouses and installed at an 8 meters height. There were four meteorological variables of interest for this research: Air temperature [$^{\circ}$], Relative humidity [%], Solar radiation [$W \cdot m^{-2} \cdot s^{-1}$] and Wind velocity [$m \cdot s^{-1}$].

3.1.3 *Phytomonitoring System*

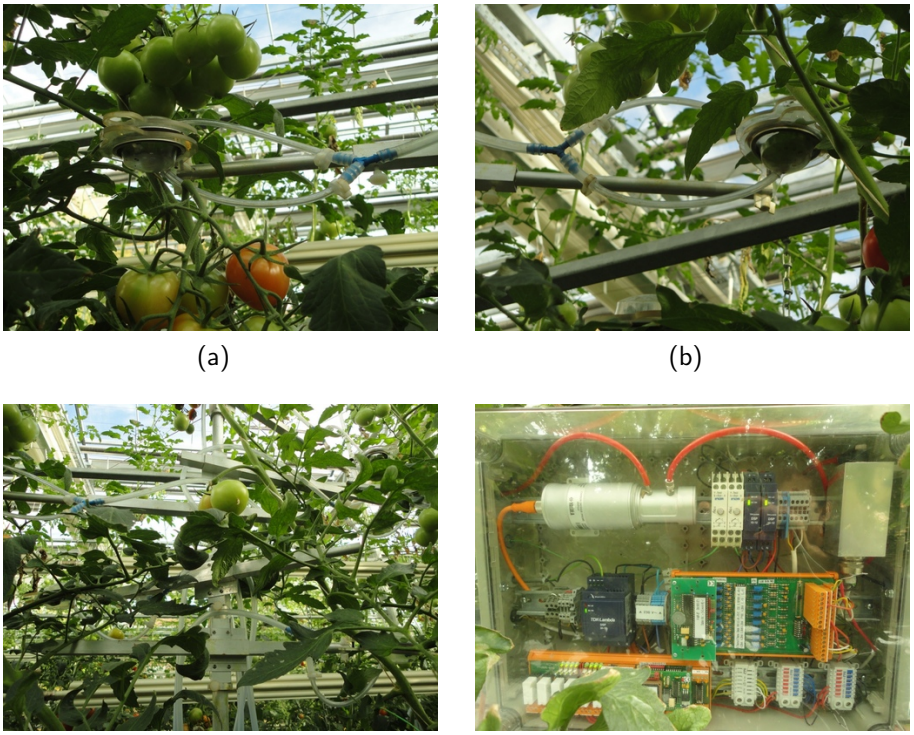
Each greenhouse was equipped with a system for measuring the three following phytometric signals:

- Leaf temperature [$^{\circ}C$]
- Transpiration rate [$mg \cdot s^{-1} \cdot m^{-2}$]

- Photosynthesis rate $[\mu g \cdot s^{-1} \cdot m^{-2}]$

The BERMONIS phytomonitoring system (BERMONIS, Steinbeis GmbH & Co. KG for Technology Transfer; Berlin, Germany) consisted on a series of plastic (PET, Polyethylenterephthalat) cuvettes that could be attached to the plants leaves without harming them (Figures 7a and 7b). These cuvettes permitted the taking of samples from the air in direct contact with the leaves, i.e. from the boundary layer next to the surface of the leaves. A small air flow was sucked out from the leaf cuvettes, in a distance of 10mm from the leaf surface, and transferred into a Dewar flask (Figure 7d), where the temperature, the air humidity, and the CO_2 content were measured. The measuring system switched valves to measure the air from the leaves and from the surrounding air. The air coming from the leaves' vicinity was mixed from 10 different cuvettes, thus getting signals from adjacent plants and from young and old, as well as shaded and non-shaded leaves.

The transpiration and photosynthesis rates were calculated directly by the instrument. These calculations were based on the differences of the absolute humidity and the CO_2 content between the reference air (which was analyzed using the same measuring equipment by switching valve groups) and the air coming from the cuvettes. The leaf temperature, on the other hand, was directly measured using a thermocouple in direct contact with the leaves, as shown in Figure 8. All these measurements were recorded in a central computer using the data acquisition system described in Section 3.1.1.



(a) (b)

(c) (d)

Figure 7: Four views of the BERMONIS phytomonitoring system. (a), (b) Cuvettes attached to the plant's leaves. (c) Plastic tubes for taking air probes from different leaves. (d) Measuring box and CO_2 sensor.

(a) (b)

Figure 8: Measuring case with the sensors for air temperature and relative humidity. (a) Front view. (b) Back view, showing also a CO_2 measuring probe.

3.1.4 *Leaf Area Index*

The Leaf Area Index (LAI) is the leaf area (upper side only) per unit area of soil below and is widely used for estimating the crop water requirements (Allen et al., 1998). The LAI represents the *leafiness* of a crop (Hunt, 1990) thus giving information about the crop growing stage. The rationale behind the use of LAI in this work is that a greenhouse with a fully developed crop shows a different behaviour than one with young plants or no crop in it. This information about the plants was intended to be used as input to the simulation models.

For each cultivation period, data from 6 weekly destructive LAI measurements in each greenhouse was available (Table 3). These measurements were used to fit a sigmoid function (Medrano et al. (2005); Rouphael and Colla (2005); Scholberg et al. (2000)) in order to create a time series suitable for the models. The sigmoidal function selected is shown in Equation (3.1.4.1), where the LAI is expressed as a function of the time (days after sowing, DFS). A total of 6 functions were fitted independently using the non-linear least squares optimization method to account for both greenhouses and the three cultivation periods. The fitted functions are shown in Figure 9 and Table 4.

$$LAI = f(DFS) = \frac{a}{1 + e^{-\frac{DFS-b}{c}}} \tag{3.1.4.1}$$

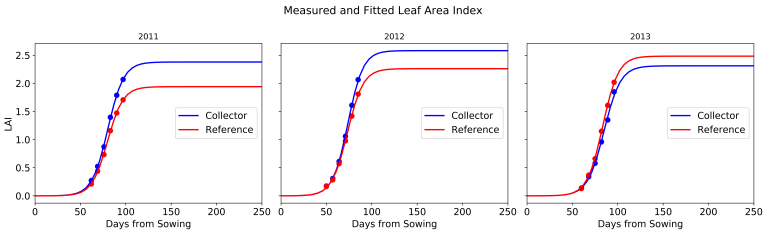


Figure 9: Measured and fitted LAI as function of the days from sowing.

	DFS	Reference	Collector
2011	62	0.21	0.27
	69	0.44	0.52
	76	0.73	0.87
	83	1.16	1.40
	90	1.47	1.79
	97	1.71	2.07
2012	50	0.16	0.18
	57	0.28	0.31
	64	0.57	0.61
	71	0.98	1.06
	78	1.42	1.61
	85	1.81	2.07
2013	60	0.13	0.14
	68	0.37	0.34
	75	0.66	0.58
	82	1.15	0.96
	89	1.61	1.35
	96	2.02	1.85

Table 3: Leaf Area Index measurements over three cultivation periods. The column DFS shows the sampling points, expressed as days after the plants were sown.

Year	Greenhouse	a	b	c
2011	Collector	2.38	80.27	8.74
	Reference	1.94	79.91	8.61
2012	Collector	2.58	73.77	8.35
	Reference	2.26	73.33	8.64
2013	Collector	2.31	84.97	9.04
	Reference	2.48	83.51	8.59

Table 4: Fitted parameters for the LAI curves (Equation (3.1.4.1)), after the measurements in Table 3.

3.1.5 *Solar Calculations*

In addition to the measurements described in previous sections, 4 theoretical variables were incorporated to the set of inputs available for the ANN models. These theoretical variables estimated to the position of the sun as a means to map the cyclical character of climate variables over time. The two main astronomical cycles which influence the short term climatic conditions are the seasons, i.e. the course of the year, and the day-night cycles.

The solar coordinates used as inputs were:

- Hour angle: ω
- Solar declination: δ
- Solar elevation: α
- Theoretical solar radiation: R_0

The hour angle ω was calculated as the difference between the current time and the time at noon (i.e. 12:00). The calculation of the other variables is given in Eq. (3.1.5.1), (3.1.5.2) and (3.1.5.3).

$$\delta = 23.45 \cdot \sin\left(\frac{360^\circ}{365.25} \cdot n\right) \tag{3.1.5.1}$$

$$\sin\alpha = \sin\varphi \cdot \sin\delta + \cos\varphi \cdot \cos\delta \cdot \cos\omega \tag{3.1.5.2}$$

$$R_0 = G_{sc} \cdot \left(\frac{\overline{d}}{d}\right)^2 \cdot \cos(90^\circ - \alpha) \tag{3.1.5.3}$$

In Eq. (3.1.5.1), n is the number of days until (or from) the spring equinox. In Eq. (3.1.5.2): $\varphi = 52.46^\circ$ (latitude the location of the greenhouses). In Eq. (3.1.5.3), G_{sc} represents the solar constant, and the terms d and $\overline{d}$ represent the current and average distance between the Earth and the Sun, respectively.

Very similar climatic conditions can be present at different times of the day or the year, yet leading to very different conditions in the near future. The estimated solar coordinates were included as a means to distinguish such situations with little computational or electronic costs.

3.2 DATASET CONSTRUCTION

The performance of mathematical models strongly depends upon the quality of the data used for building them. This section deals with the preparation of the data before the model construction started. The data preparation was performed in two subsequent processes.

The first process included checking for data integrity and erroneous values, but also sycnhronisation of time steps and smoothing. The assertion of data quality is a major concern in model building in general, and in the training of ANN in particular. The calculations and adjustments made to the data prior to model building are described in Section 3.2.1.

After the aforementioned process was finished, the data was divided into data groups, thus forming the actual *datasets* that were used to shape the model later on. Although the concrete way of integrating a dataset, i.e. dividing the complete amount data into groups, might vary according to the type of model (and design methodology), it is a common practice (Ripley, 2006; James et al., 2013) to consider at least two different datasets: For adjusting the model parameters (*training*) and for model evaluation after these parameters were set (*test*). In this work, a third dataset was used, to monitor the training process itself (*validation*). The construction of the datasets is described in Section 3.2.2.

Regarding the dataset construction, the following particular concerns were present in this work:

- The data must permit an interpretation in *time*, in order to construct predictions of climate.
- The data must refer to the actual cultivation periods.
- The *test* dataset must consist (at least partially) of a series of adjacent records to facilitate a graphical interpretation.

- The dataset must allow to build models from each greenhouse separately.

3.2.1 *Data Preprocessing*

The ZINEG greenhouse facility provided a year-round series of climate and weather measurements. For this research, only those listed in the first column of Table 6 were used. Despite the year-round disponibility of measurements, the periods of useful data were limited to those with actual phytometric measurements (see Table 5 and Figure 10). Limiting the dataset to these periods excluded the winter pauses as well as the periods when the plants were very young. The valid records from all three cultivation periods made up a total of 380,590 data points (190,295 for each greenhouse), calculated every 5 minutes.

This fixed time step came from the the Plantputer software (Section 3.1.1), that delivers a single value of every variable over this interval. Having the values over a 5-Minutes period presented advantages particularly well-suited for this project:

1. It synchronised the measurements from different sources, which were otherwise recorded with different time stamps.
2. It diminished the presence of oscillations, thus acting as a low-pass filter.
3. It set the time steps to a scale suitable to make discrete predictions of climate.

Three more preprocessing steps were applied to the series of data before their division into datasets:

- Search for untypical values. Those values falling outside the range shown in Table 6 were considered outliers and were consequently replaced by the average of the previous and next measurements.
- Low-pass filtering. A Savitzky-Golay filter (window size: 35; polynomial order: 3) was applied to diminish high frecuency noise. This treatment was not applied to those variables that normally change their value rapidily: Ventilation opening, Thermal screen closure, Heating energy and Cooling energy.

- The tables were reshaped to include 5 complete records in each line. This was necessary because the time references would be lost in the randomization process described in the next section. Each line included thus 5 time steps, which could later on be interpreted as 3 *past*, 1 *present* and 1 *future* measurements.

Cultivation period	2011	2012	2013
Sowing	11/01/2011	05/12/2011	20/12/2012
Planting	14/03/2011	24/01/2012	18/02/2013
Beginning of model	02/05/2011	11/02/2012	22/02/2013
End of model	24/11/2011	25/09/2012	12/11/2013

Table 5: Cultivation periods. The rows *Beginning* and *End of the model* indicate the periods available for model building. See also Figure 10.

	Variable	Min	Max	Units
Climate	Air Temperature	10	50	$^{\circ}C$
	Relative Huimidity	20	100	%
	CO_2 Concentration	50	1500	ppm
Actuators	Heating power	-100	300	kW
	Cooling power	-300	100	kW
	Ventilation opening	0	100	%
	Thermal screen closure	0	100	%
Weather	Air Temperature	-20	50	$^{\circ}C$
	Relative Humidity	5	100	%
	Solar Radiation	-10	1500	$W \cdot m^{-2}$
	Wind Velocity	0	30	$m \cdot s^{-1}$
Phytometric signals	Leaf Temperature	10	35	$^{\circ}C$
	Transpiration Rate	0	150	$mg \cdot s^{-1} \cdot m^{-2}$
	Photosynthesis Rate	-10	30	$\mu g \cdot s^{-1} \cdot m^{-2}$

Table 6: Range of valid measurements for identifying outliers.

3.2.2 *Dataset Construction*

After preprocessing, 379,336 valid data records (189,668 from each greenhouse) were available to integrate the *training*, *validation* and *test* datasets needed for building the models. This approach for model comparison and selection is called the *hold out method* (Bishop, 1995).

The first step was to set aside the *test* dataset. The *test* dataset included 104 series of 288 adjacent records (each representing a calendar day), evenly distributed over the last cultivation period (see Figure 10). This dataset was excluded from the model construction steps and only used to evaluate the performance of the models, after their training was over. The fact that they were chosen as adjacent records, prevented them from losing reference to their original place in the dataset if they were to be randomly picked. This was important for two reasons: First, it allowed to compare the model performance in each greenhouse separately; and second, it permitted to run simulations suitable for graphical representation on time.

The rest of the records were then randomly divided into training and validation datasets. All these records were normalized in the range [-1,1] according to Eq. (3.2.2.1). The maximum and minimum values used were those of their combined set. These values were stored to scale the simulations of the *test* dataset.

$$x_{norm} = 2 \cdot \frac{x - min(x)}{max(x) - min(x)} - 1 \tag{3.2.2.1}$$

Three versions of the *train* and a *validation* datasets were integrated: One with data from a each single greenhouse and a third one including records from both. The number of records used in each version is given in Table 7 and a diagram depicting the integration of the dataset is given in Figure 11.

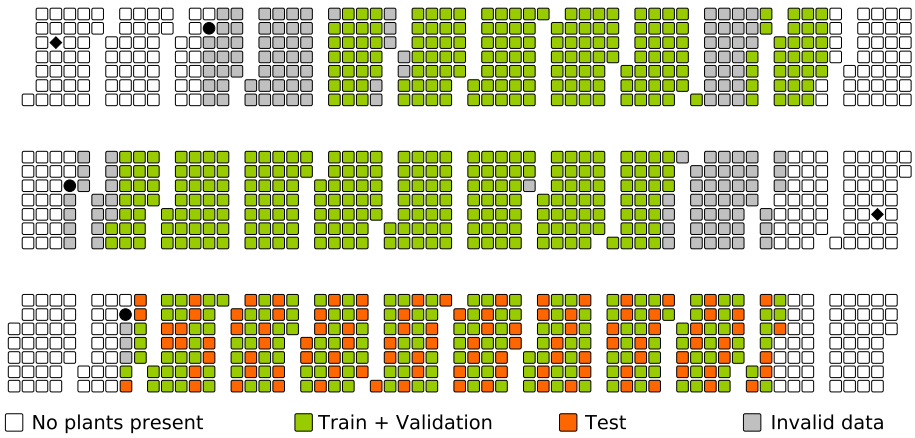


Figure 10: Calendar showing selected dates for the three cultivation periods. From top to bottom: Years 2011, 2012 and 2013. From left to right: Months from January to December. Diamonds tag the sowing dates and circles the transplantation dates (the point where the plants came in the production greenhouses). The areas coloured in green indicate the periods with valid phytometric signals used to train the models. The areas coloured in orange indicate the days taken for the *test* dataset. The areas coloured in gray indicate periods with non-valid phytometric measurements.

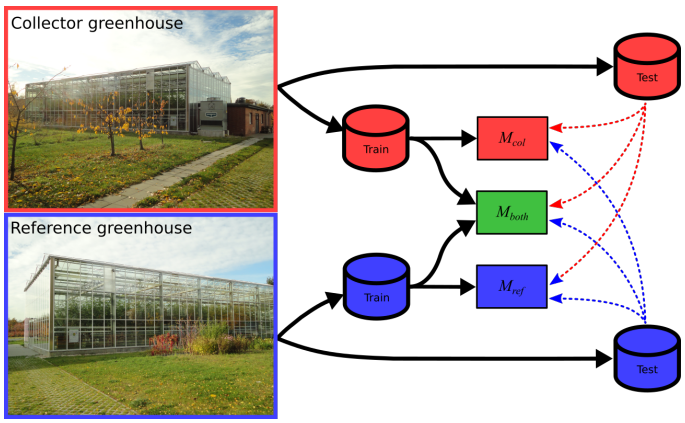


Figure 11: Dataset construction, depicting the origin of the data and their use. The symbols labeled *train* include both the train and validation datasets. The rectangles show the models created related to the data origin for each. The models are described in detail in Section 3.3.

Dataset	Data origin	Number of records
Training	Collector	136,289
	Reference	136,289
	Both	272,578
Validation	Collector	24,051
	Reference	24,051
	Both	48,102
Test	Collector	29,951
	Reference	29,951
Total		379,336

Table 7: Dataset division. The second column shows the greenhouse where the data for each dataset version was collected. The data was never mixed for the *test* dataset.

3.3 MODEL CONSTRUCTION

This section describes the design and implementation of the simulation models in detail. It is organized in three parts, the first telling about the single, standalone models and their architecture, the second about the inputs used in each model, and the third about the software implementation.

There are two standalone models described in this work:

- **Climate Prediction Model (CM).** Aimed at the prediction of the air temperature and relative humidity inside a greenhouse 5 minutes ahead in time (i.e. a *One-Step Prediction (OSP)* of the inside climate).
- **Phytometric Signals Model (PM).** Aimed at the estimation of leaf tissue temperature, transpiration and photosynthesis rates.

The nature of the data available (Table 7) allowed to build three variations of each model (see also Figure 11):

- Trained with data from the first greenhouse (collector)
- Trained with data from the second greenhouse (reference)
- Trained with data from both greenhouses

The following sections describe the variations and tests carried out in the model design phase, which involved model comparison during the validation dataset. These comparisons helped to take the decision upon the form that a final model would take. After this design phase, three final models (for each greenhouse, and both) were trained and used for simulations of the test dataset.

3.3.1 *Architecture of the models*

Design considerations on model structure

The first model had the following constraints:

- It had to deal with a time-representation model to be able to generate forecasts of the variables. To permit this behaviour, the dataset had to be adjusted as described in Section 3.2.2.
- The air temperature and relative humidity are closely intertwined, which suggested that a single network with two outputs could be an suitable model.

The structure of the second model was different. This model did not have to cope with the time-representation problem of the climate model, but had the following considerations:

- The leaf temperature can often be simulated with climatic information, since it closely depends on the air temperature.
- The simulation of the transpiration rate could be improved by taking the leaf temperature as an input.
- The simulation of the photosynthesis rate (by all means a very complicated physiological process) could be improved by including the leaf temperature and transpiration rate as inputs.

The aforementioned premises allowed to set up a composite model, i.e. a cascade configuration where the simulation of a signal helped improve the next ones. For the case of ANN, this implied the design and implementation of 3 independent models (three independently trained ANN) which can later on be connected together (see Figure 13b).

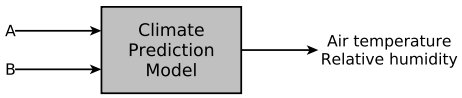


Figure 12: Climate prediction model. The input variables can be seen in Table Table 8.

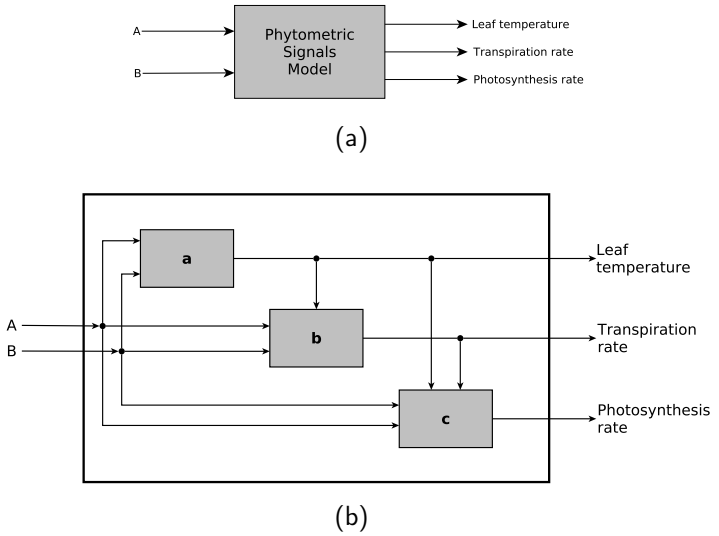


Figure 13: Phytometric signals model. (a) Overview. (b) Cascade configuration, showing the independent ANN models for each phytometric signal. See Table 8 for the specific input variables.

ANN architecture

The models shown in Figures 12 and 13 were implemented using ANN in the form of multilayer perceptrons. The number of hidden layers was set to one (Section 2.2) and three possibilities were explored regarding the number of hidden nodes:

- ANN_1: $n_h = \sqrt{n_i \cdot n_o}$
- ANN_2: $n_h = n_i$
- ANN_3: $n_h = 2 \cdot n_i + 1$

In these formulae, n_i denotes the number of input nodes, n_h the number of hidden nodes and n_o the number of output nodes in each particular ANN. While the value of n_o is fixed ($n_o = 2$ for the climate model and $n_o = 1$ for each ANN in PM), the number of inputs (i.e. the value of n_i) was yet to be defined. In order to do this, a number of inputs were tested, as described in Section 3.3.2. The size of three different hidden layers can be graphically appreciated in Figure 14, with

an example of an ANN with 8 input and 2 output nodes (the simplest configuration tested for CM, see Table 8).

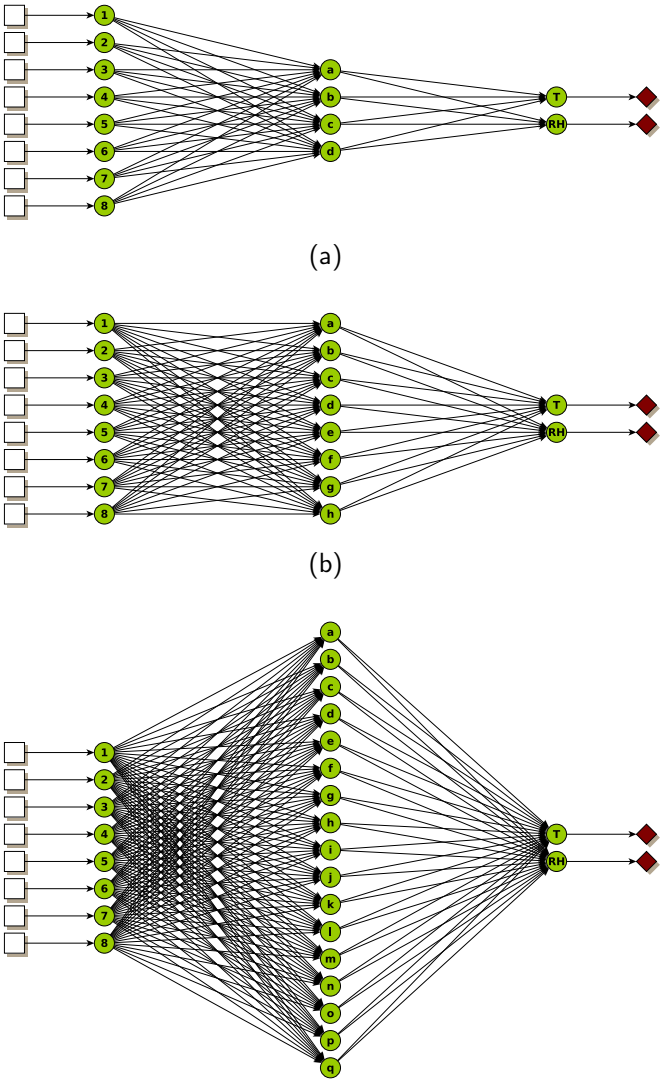


Figure 14: Three ANN with different number of hidden nodes.

3.3.2 *Inputs & Outputs*

Each model was built using two different sets of input variables: The first set (A) was constituted by a series of inputs which represented the starting point for the design, which means that the variables in A represented the smallest set of variables to be included as inputs in the final model. The second set (B) included those variables which needed to be further investigated to be eventually included in the model. In other words, the set B included additional variables, which could also be used as inputs to to the model. This was be defined as:

$$\begin{aligned} A =& \{I_i \mid 0 < i \leq m\} \\ B =& \{I_i \mid m < i \leq n\} \\ C =& \{I_i \mid 0 < i \leq n\} \\ =& A \cup B \end{aligned} \tag{3.3.2.1}$$

where a third set C was included as the set containing all available inputs and I_i represents a single input variable. Also $A \subset B \subset C$.

If $M(min)$ denotes the model created with the inputs in A, and $M(all)$ the model with all the inputs in C, then two models can be tested for each variable in B:

- $M(min + i)$: Model with inputs $\{A + I_i\}$
- $M(all - i)$: Model with inputs $\{C - I_i\}$

The variables used as inputs are listed in Table 8, along with the corresponding set they are assigned to. It must be noted that because the ANNs were built independently from each other, the input sets were different for each of them.

Time-lagged inputs in the climate model

Additionally, the inputs to the climate model were tested in 4 different time lags:

- $[t]$
- $[t, t-1]$
- $[t, t-1, t-2]$
- $[t, t-1, t-2, t-3]$

where the numbers give the relative time step to which the variable was bound, thus t refer to current values (at the point when each simulation was run). Since a time step had a duration of 5 minutes (Sections 3.1.2 and 3.2.1), $t-1$, $t-2$ and $t-3$ refer to values 5, 10 or 15 minutes before (in the past), respectively.

Cascaded inputs in the phytometric signals model

The inputs for PM had a structure that allowed the connection of the internal submodels, as was previously shown in Figure 13. All inputs in the set A were valid for the three ANN, i.e. the three submodels shared the inputs in A. However, the inputs in set B were different for the three models because they included variables generated internally: The output of PMa served as input to PMb and PMc, and the output from PMb served as input to PMc.

Variable	CM	PM		
		a	b	c
Air Temperature	$I_1 \in A$	$I_1 \in A$	$I_1 \in A$	$I_1 \in A$
Relative Huimidity	$I_2 \in A$	$I_2 \in A$	$I_2 \in A$	$I_2 \in A$
Ventilation opening	$I_3 \in A$	–	–	–
Thermal screen closure	$I_4 \in A$	–	–	–
Hour Angle	$I_5 \in A$	$I_3 \in A$	$I_3 \in A$	$I_3 \in A$
Solar Declination	$I_6 \in A$	$I_4 \in A$	$I_4 \in A$	$I_4 \in A$
Solar Elevation	$I_7 \in A$	$I_5 \in A$	$I_5 \in A$	$I_5 \in A$
Theoretical Solar Radiation	$I_8 \in A$	$I_6 \in A$	$I_6 \in A$	$I_6 \in A$
Heating power	$I_9 \in B$	–	–	–
Cooling power	$I_{10} \in B$	–	–	–
CO_2 Concentration	–	$I_7 \in B$	$I_7 \in B$	$I_7 \in B$
Leaf Temperature	–	–	$I_8 \in B$	$I_8 \in B$
Transpiration Rate	–	–	–	$I_9 \in B$
Leaf Area Index	$I_{11} \in B$	$I_8 \in B$	$I_9 \in B$	$I_{10} \in B$
Air Temperature outside	$I_{12} \in B$	$I_9 \in B$	$I_{10} \in B$	$I_{11} \in B$
Relative Humidity outside	$I_{13} \in B$	$I_{10} \in B$	$I_{11} \in B$	$I_{12} \in B$
Solar Radiation outside	$I_{14} \in B$	$I_{11} \in B$	$I_{12} \in B$	$I_{13} \in B$
Wind Velocity outside	$I_{15} \in B$	$I_{12} \in B$	$I_{13} \in B$	$I_{14} \in B$

Table 8: Inputs available for model building. The letters A and B indicate the group where that particular variable belongs to. Note that the sets A and B were different for each model, and are shown in this table together only for convenience.

3.3.3 *Model implementation*

The FANN 2.2.0 library (Fast Artificial Neural Networks: `http://leenissen.dk/fann/wp/`) was selected for the implementation of the ANN. This library had the advantage of being released under the GNU Lesser General Public License. The networks were coded in the C Programming Language.

The ANN were constructed as multilayer perceptrons, with a symetric sigmoid activation function (Eq. (3.3.3.1)) and a bias unit for the hidden and output layers. The weights were initialised using the Nguyen-and-Widrow algorithm, run on the corresponding training dataset (Nguyen and Widrow (1990a)). This algorithm set the network's weights to random values as a way to improve the training process, both in speed and convergence. A side-effect of this random initialisation is that the same network architecture may converge to different local minima of the error each time a learning process is carried out.

The learning process (network training) was performed using the normalized inputs and outputs (Section 3.2.2). The learning algorithm used was the resilient backpropagation (Riedmiller and Braun, 1993) (a variation of the original backpropagation algorithm), with the mean squared error (MSE) as cost function. During the design phase, all ANN were trained for 500 epochs, during which the MSE on both the training and test dataset were recorded. These values were used to plot the learning curves used to verify the training process after its completion.

Additionally, tThe MSE after 500 training epochs was stored for the $M(min)$ and $M(all)$ versions of each model (Section 3.3.2) in a different fashion. These models were initialized 15 times to have a measure of the inherent random variation of each ANN architecture. This measurement helped to select the final models.

The final models were trained for a total of 10,000 epochs. The learning curves were examined to search for pathological patterns, and the full-trained ANN were then run over the test dataset to calculate the MSE and to plot run charts for performance evaluation.

$$\begin{aligned} y &= \tanh(x) = \frac{2}{(1+e^{-2\cdot x})} - 1 \\ \frac{dy}{dx} &= 1 - y^2 \end{aligned} \tag{3.3.3.1}$$

4

RESULTS

4.1 ANN FOR GREENHOUSE CLIMATE PREDICTION

4.1.1 *Model Design*

The combination of 3 numbers of hidden nodes and 4 time lags, when applied on the models $M(min)$ and $M(all)$ (see Section 3.3.2 on page 59) made a total of 24 networks, which were trained on 3 different datasets (Table 7). Each of these networks was initialized and trained 15 times, yielding different validation errors (MSE). These results can be seen on Figure 15.

The graphs in Figure 15 show a recurrent pattern, in a $\vee$-shaped curve, that points out a better performance (lower error) of the network ANN_2. Indeed, the mean error was lower in 20 out of 24 tests (all in case of the dataset from the reference greenhouse).

An analysis of variance (ANOVA) confirmed the MSE for ANN_2 to be significantly smaller than those of ANN_1 and ANN_3 for both datasets with single greenhouse data. On the combined dataset the ANN_2 performed significantly better than ANN_3, but only slightly better than ANN_1 (Table 9).

The networks with as many hidden nodes as inputs ($n_h = n_i$) had an overall better performance for climate prediction. Thus, the network design continued with a closer look at the number of previous measurements with better behaviour under ANN_2.

It can be seen in Figure 16 that the model $M(min)$ (with a minimum of inputs) in all cases outperformed $M(all)$ when a time lag of 1 or 2 waxs added. The same diagram shows that the validation error increased when 3 previous steps of data were used as inputs. While an ANOVA test showed only partially significant differences between the

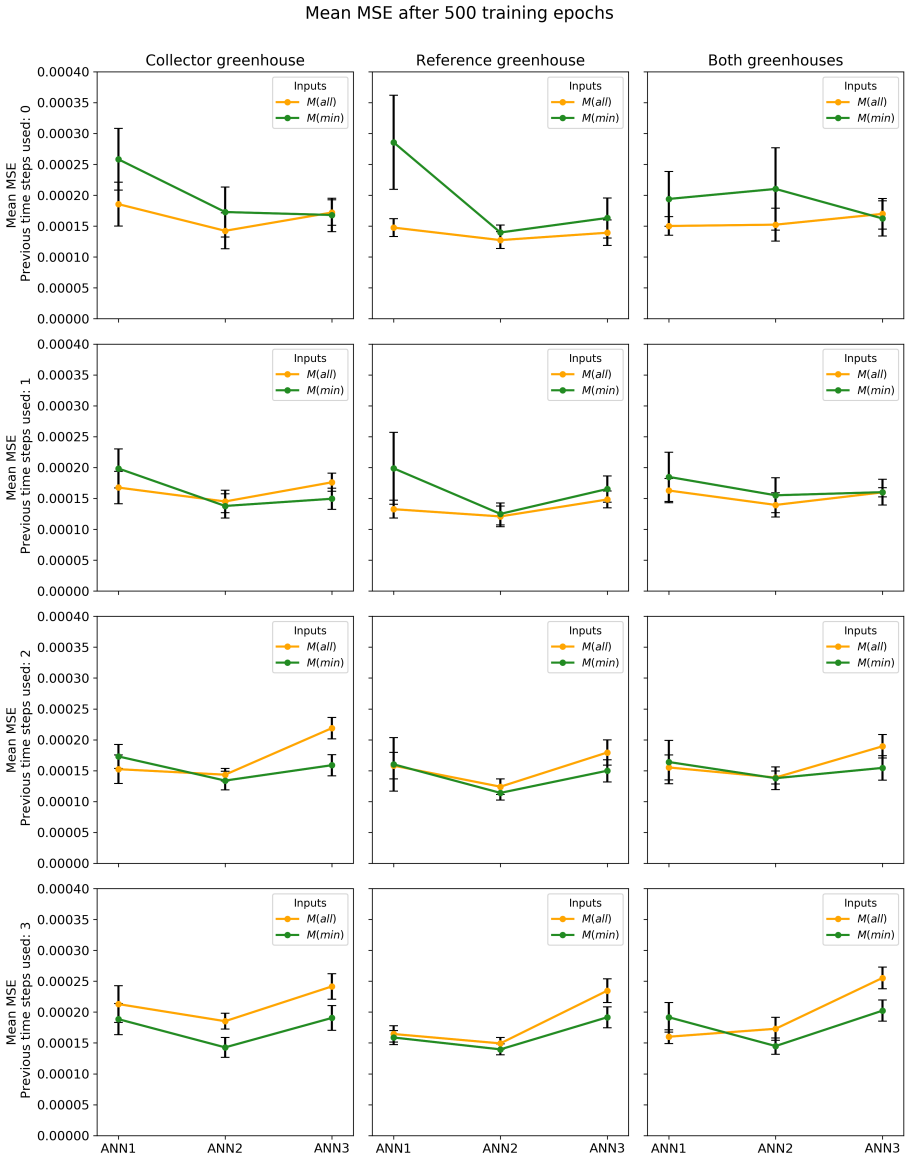


Figure 15: Climate prediction model: MSE achieved after 500 training epochs, using different network architectures and datasets. The points indicate the mean of 15 MSE values (independent trainings). The error bars show the 95% confidence interval of the mean.

Training Dataset	Architecture of the hidden layer		
	ANN_1	ANN_2	ANN_3
Collector	0.000192 (a)	0.000150 (b)	0.000184 (a)
Reference	0.000176 (a)	0.000130 (b)	0.000171 (a)
Both	0.000170 (ab)	0.000156 (b)	0.000182 (a)

Table 9: Climate prediction model: Mean validation MSE and groups calculated over the number of hidden neurons. Every dataset was analysed independently (Tukey's Honest Significant Difference test, $p < 0.0005$).

number of previous steps used, the best results were obtained with a data lag of 2 steps.

The selected architecture (ANN_2 with 2 previous steps of data) was then trained with the inputs shown in Table 8. The results of the training process of the models $M(min + i)$ and $M(all - i)$ (as defined in Section 3.3.2) is shown in Figure 17.

The most striking feature on Figure 17 is the overlapping of the confidence intevals for $M(min)$ and $M(all)$ in the three models. This suggested that this model architecture could not perform better when more input signals were incorporated. Moreover, the MSE for $M(all)$ was bigger in all cases shown. A t-test confirmed no differences in the means for $M(min)$ and $M(all)$, in any of the three datasets ($p > 0.2$ in all cases).

The direction and length of the arrows in Figure 17 give insights on the behaviour of each individual input variable. A long yellow arrow pointing downwards indicates a variable suitable for selection on a *forward* approach. Similarly, a long green arrow pointing upwards indicates a variable suitable for selection on a *backward* approach. A concurrence of these two cases was taken as indicative of the pertinence to include a variable in the final model.

The final models were set to include only the minimum inputs set because no clear patterns for input selection could be identified in Figure 17.

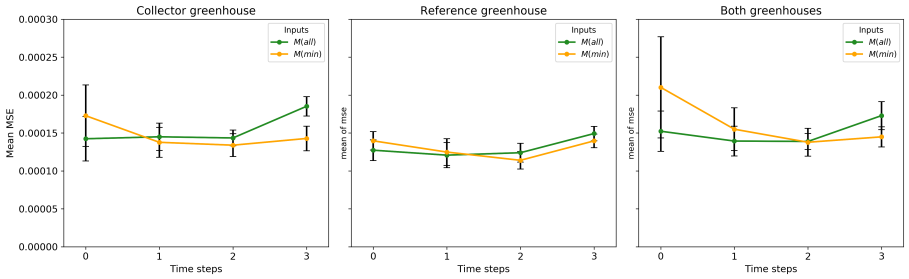


Figure 16: Climate prediction model: MSE achieved after 500 training epochs, using $n_h = n_i$ hidden nodes and three different datasets. The points indicate the mean of 15 MSE values (independent trainings). The error bars show the 95% confidence interval of the mean.

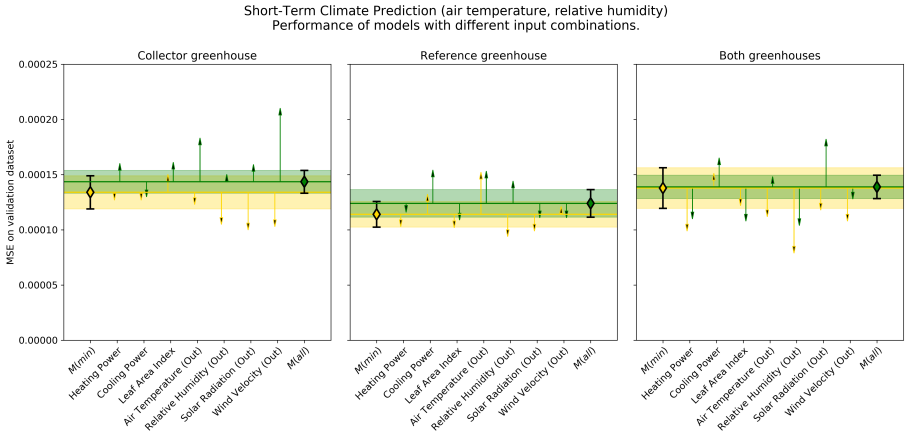


Figure 17: The selected network architecture (ANN_2) for the climate prediction model, trained with different combinations of inputs and two previous steps of data. The error bars and coloured stripes show the 95% confidence interval of the mean for the models $M(min)$ (yellow) and $M(all)$ (green). The yellow arrows show the error of a model incorporating the corresponding input. The green arrows show the error of a model lacking the same input.

The selected model design is summarised in Table 10. The diagrams in Figure 18 show the evolution of the training and validation error during the 10,000 training epochs. These diagrams show no signs of overfitting or other problems during training.

	Selected parameters
Hidden layer nodes	ANN_2: $n_h = n_i = 24$
Time steps included	$[t, t-1, t-2]$
Input signals	Air temperature Relative humidity Ventilation opening Thermal screen closure Hour angle Solar declination Solar elevation Theoretical solar radiation

Table 10: Climate prediction model: Summary of model design.

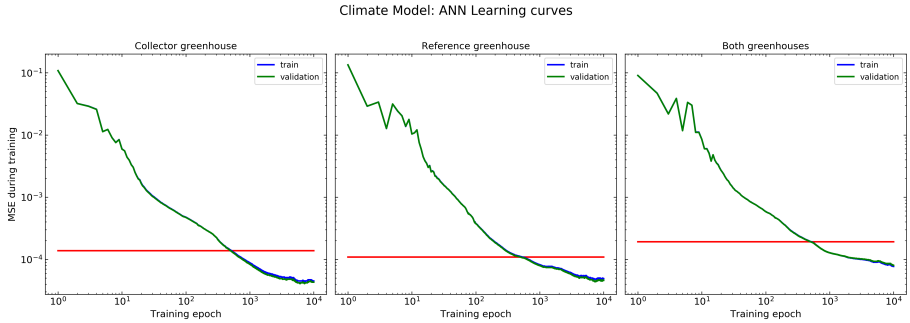


Figure 18: Climate prediction model: Training of the final models over 10,000 epochs. After the first 10 training epochs (approximately), the error follows a smooth decreasing path. The red lines show the MSE at 500 training epochs (for the corresponding model) and are included for reference only.

4.1.2 *Model Test*

Overall Results

The one-step prediction for both climate outputs gave very good simulation results over the whole validation dataset, regardless of the training dataset used . However, there were also differences in performance, not only between training datasets, but also between air temperature and relative humidity, as well as between test dataset and different points in time.

In Figures 19 and 20 we can see scatterplots that show that the simulation results were not evenly distributed along the 45° line: They deviated from this line on both ends, trimming both the high and the low values of both variables. The restriction is more noticeable by lower temperatures and higher relative humidities.

The errors on both ends of the scatterplots were constrained to a range in the y-axis,suggesting that the ANN saturated (or got close to saturation). A possible explanation was given by the sigmoid activation function used (Eq. (3.3.3.1)).

The predicted temperature showed bigger errors when the actual (measured) temperature was low (below 15°). This occurred on both test greenhouses, but not with the three training datasets: The collector-trained ANN behaved slightly better at lower temperatures. Nevertheless, these errors represented only a very small fraction of the whole dataset (6 events over the year).

On the other hand, the biggest errors on relative humidity presented themselves at higher values, when results of near 100% were expected, but simulated values around 90% were obtained instead. The upper end of Figure 20 also makes clear that the networks did not give results above 100% relative humidity at all.

OB The model trained on data from the collector greenhouse yielded lower errors on relative humidity. This can be appreciated at the individual points further away from the 45° in Figure 20, specially those located in the lower half of the diagram. These values will be analised in detail on the next section (Section 4.1.2 on page 75).

The distribution of the absolute errors is presented in Figures 21 and 22. In these figures, the top curves compare the errors by means of kernel density estimation (KDE) curves. The curves were computed with gaussian kernels using the *scipy* software library.

The boxplots in the bottom panels give another view at the error distribution. The boxplot whiskers indicate the percentiles 1 and 99, thus encompassing 98% of all test points. Those errors falling further from the whiskers were omitted in these diagrams, for the sake of visual simplicity. The next section (Section 4.1.2) deals with these particular cases in detail.

The most striking result featured in Figure 21 is that the temperature in the reference greenhouse could be better estimated with the model trained in the collector greenhouse. This can be seen at the lower spread of the KDE-curve, as well as the relative sizes of the boxplots. The opposite situation did not hold: The model CM_{ref} showed a strong positive bias and higher dispersion when tested on the collector greenhouse.

Also the model CM_{both} predicted the air temperature in the reference greenhouse better, largely outperforming both single-greenhouse-trained models. Although its overall performance on the collector greenhouse was seemingly good as well (Figure 21), a further analysis of the simulations revealed patterns (on both test greenhouses) that cannot be seen in this overall view. These patterns will be presented in Section 4.1.2.

The model CM_{col} predicted the relative humidity with more accuracy, with less dispersion of the errors around zero as well as a more symmetric distribution (Figure 22), regardless of the test greenhouse. Only when the relative humidity fell below 40% the CM_{col} model delivered wrong results. For lower relative humidity values, the model CM_{both} performed better in both greenhouses.

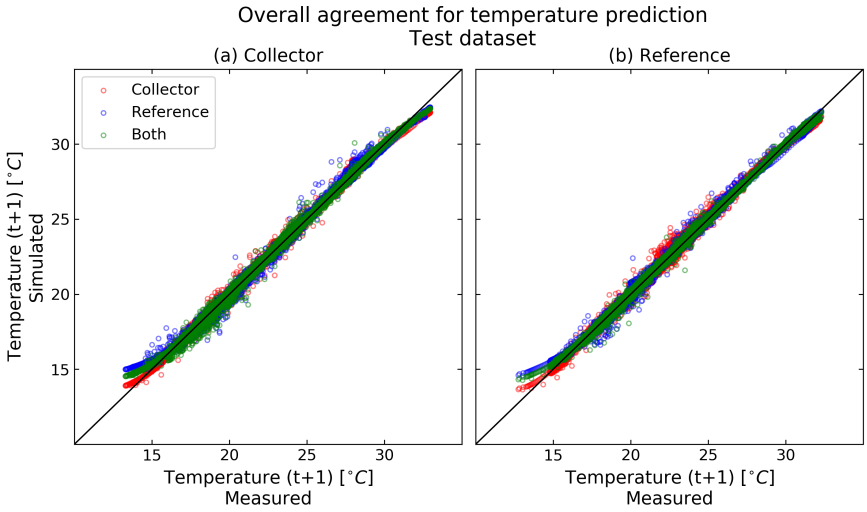


Figure 19: Measured and simulated temperature (one-step prediction). Left: Collector greenhouse. Right: Reference greenhouse. The colours refer to the dataset used for training.

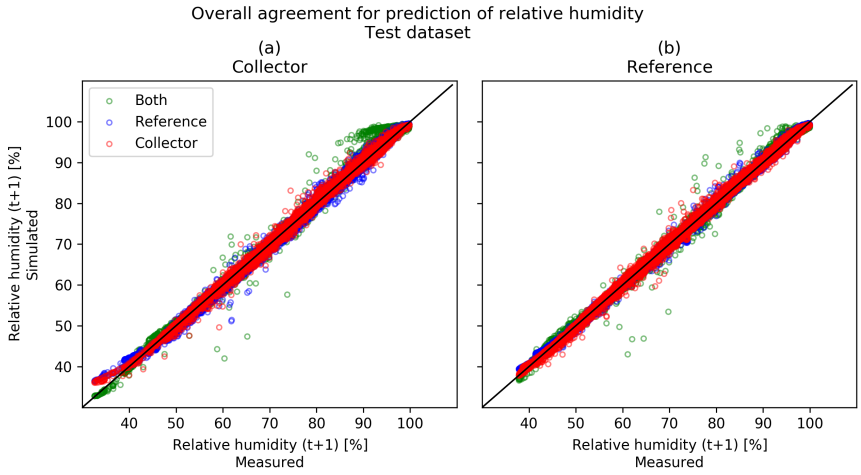


Figure 20: Measured and simulated relative humidity (one-step prediction). Left: Collector greenhouse. Right: Reference greenhouse. The colours refer to the dataset used for training.

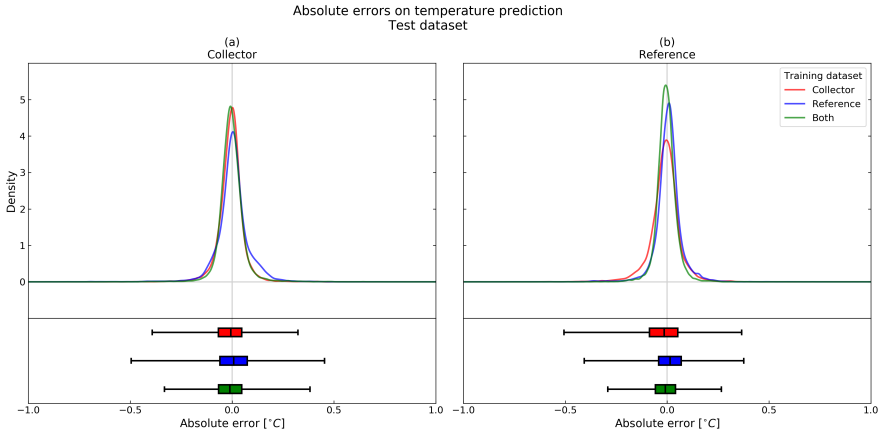
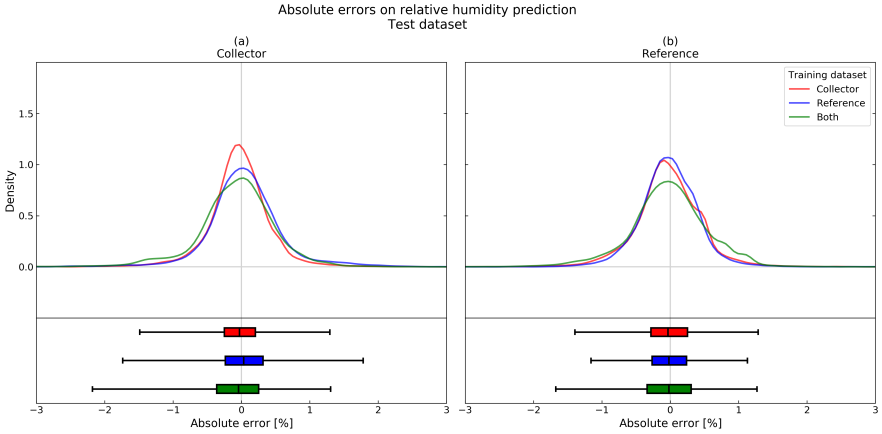


Figure 21: Absolute errors on one-step prediction of air temperature. Top: Kernel density estimation of the errors. Bottom: Distribution of the same errors. The whiskers mark the percentiles [1,99]. Left: Collector greenhouse. Right: Reference greenhouse. The colours refer to the dataset used for training.



t

Figure 22: Absolute errors on one-step prediction of relative humidity. Top: Kernel density estimation of the errors. Bottom: Distribution of the same errors. The whiskers mark the percentiles [1,99]. Left: Collector greenhouse. Right: Reference greenhouse. The colours refer to the dataset used for training.

Selected Test Periods

The selected examples shown in this section incorporate the temporal relationships between the individual simulations, which were not taken into account in the previous analysis. When the measurements and simulations were represented along the time axis, it was possible to see the effect of previous input values on the results. It was also possible to identify in which cases the simulations presented bigger errors and how these errors related in sequences.

Specifically, the following situations of interest will be presented:

- The most common scenario, i.e. smooth simulation curves with small deviations from the measurements
- Single error situations were a simulation resulted in a single strong deviation (peaks)
- Sustained deviations lasting several time steps (drifts)

The aforementioned situations will be illustrated with the aid of 2 test periods, each of which comprises 2 days of data (as described in Section 3.2.2). They were selected because the different effects can be clearly seen in different parts of the run charts. These charts incorporate the state value of the 2 actuators used as inputs to the model: the percent opening of the roof ventilation and the percent closure of the thermal screen above the canopy. It will be shown that the changes in the states of the actuators strongly influenced the simulation results.

It is worth mentioning that the change of state of the actuators took place very rapidly: both the ventilation and the thermal screen were able to change their opening value in 100% withih the 5-minutes measuring interval.

The comparatively big errors marked with a letter **A** in Figures 23 and 24 coincided with changes in the ventilation opening and were more intense in the model CM_{col}. Also, the errors in the simulation of temperature were more striking than those in the simulation of relative humidity.

On the other hand, the rapidly changing values of the thermal screen closure increased the errors from both CM_{ref} and CM_{both}. These errors are marked with a letter **B** in Figures 23 and 24, and occurred only once

in the test period shown. Note that a single change in the thermal screen produced a single peak in the simulated signal (both for temperature and relative humidity), with the values recovering afterwards.

The period that is labeled in the figures with a letter **C** indicates a sustained drift of the signal. This kind of deviation was characterised by a simulated signal that followed the measurements' trend with a comparatively low bias.

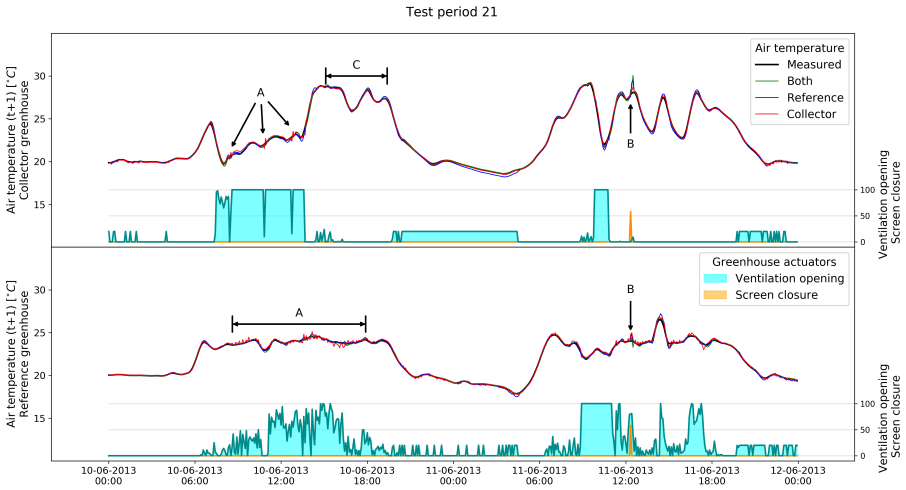


Figure 23: One-step prediction of temperature, as tested with data from the 10th and 11th of June, 2013. Top: Collector greenhouse. Bottom: Reference greenhouse. The coloured lines refer to the 3 independently trained models. The shaded areas on the right axis indicate the state of the actuators. The letters **A**, **B** and **C** tag different types and sources of error, described in the text.

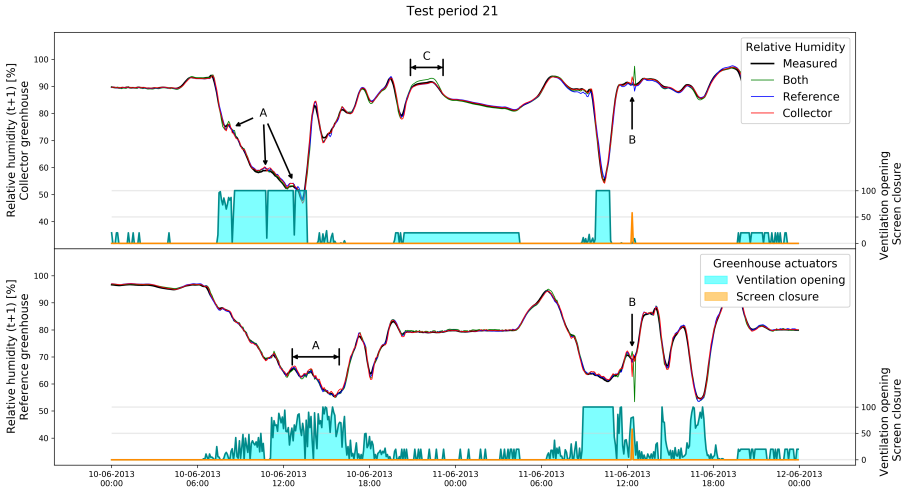


Figure 24: One-step prediction of relative humidity, as tested with data from the 10th and 11th of June, 2013. Top: Collector greenhouse. Bottom: Reference greenhouse. The coloured lines refer to the 3 independently trained models. The shaded areas on the right axis indicate the state of the actuators. The letters **A**, **B** and **C** tag different types and sources of error, described in the text.

The second selected test period is depicted in Figures 25 and 26, with simulations of air temperature and relative humidity, respectively. The letter tags used in these figures refer to the same situations previously described, namely: **A**: Errors due to changes in the ventilation opening, **B**: Errors due to changes in the thermal screen closure, and **C**: Signal drift. The letters **D** and **E** represent particular cases of model reactions to the operation of the thermal screen.

The label **D** shows a particular type of error, a very localised drift occurring when the thermal screen was only partially opened, most commonly in the morning. The error lasted only as long as the screen was partially closed, receding when it opened completely.

A similar effect (marked with the tag **E**) can be seen when the thermal screen closed in the evening. A close-up of the particular case marked with **E**$\star$ in Figure 26 is presented in Figure 27.

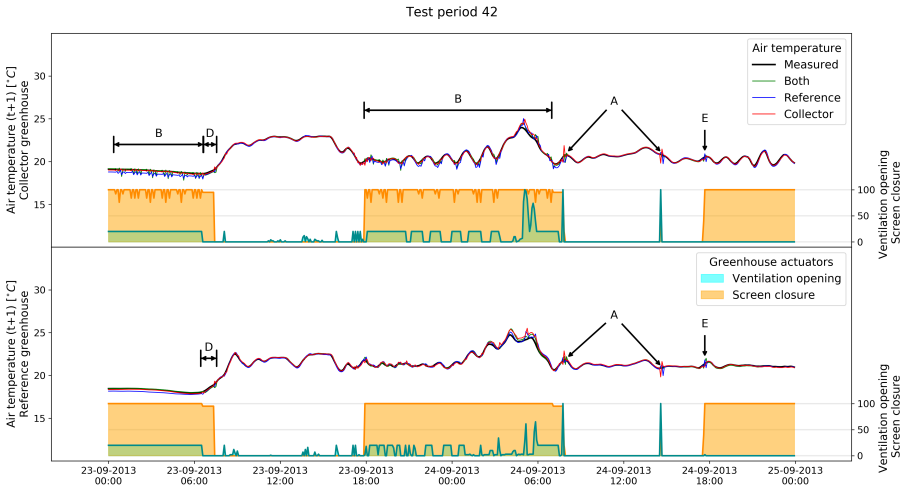


Figure 25: One-step prediction of temperature, as tested with data from the 23rd and 24th of September, 2013. Top: Collector greenhouse. Bottom: Reference greenhouse. The coloured lines refer to the 3 independently trained models. The shaded areas on the right axis indicate the state of the actuators. The letters **A**, **B**, **C**, **D** and **E** tag different types and sources of error, described in the text.

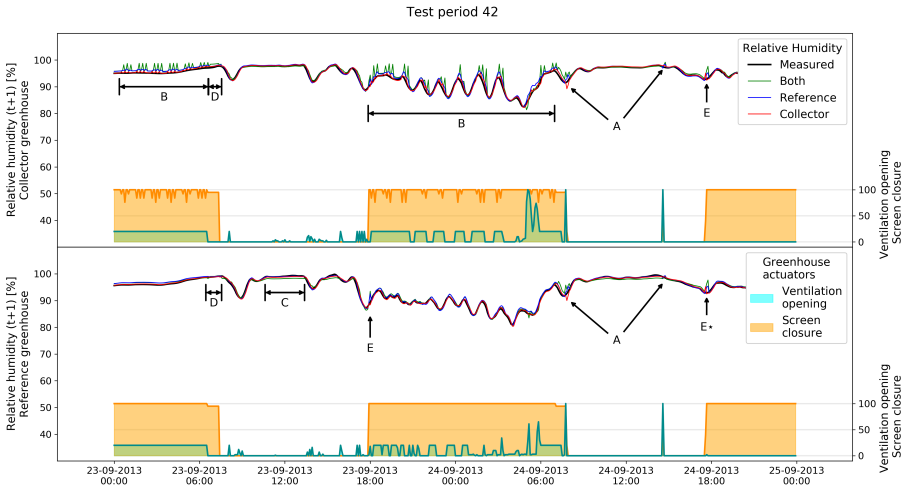


Figure 26: One-step prediction of relative humidity, as tested with data from the 23rd and 24th of September, 2013. Top: Collector greenhouse. Bottom: Reference greenhouse. The coloured lines refer to the 3 independently trained models. The shaded areas on the right axis indicate the state of the actuators. The letters **A**, **B**, **C**, **D** and **E** tag different types and sources of error, described in the text.

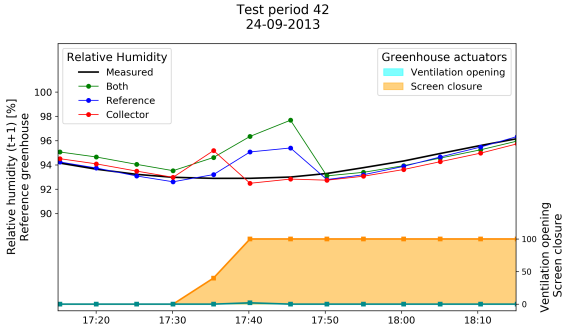


Figure 27: One-step prediction of relative humidity, showing the time in the evening when the thermal screen closes (orange surface). The roof ventilation remained closed during the time depicted. Values from the reference greenhouse, on September 24th, 2013.

The graph in Figure 27 shows the closure of the thermal screen before the night, along with the triggered peak errors in all models. This example focuses on a single event (at the reference greenhouse) for clarity.

Since the models took the input variables in three time steps (t, $t-1$, $t-2$), they seemed to overreact to the fast actuator changes only during this time window. In Figure 27, the error started to increase at 17:30, when the current ventilation opening (t) was 40% (and the two previous still 0%). Notice that in this case, the screen was fully closed within 10 minutes. The simulations regained their performance by 17:50, i.e. exactly 2 time steps after the screen was completely closed.

Although the example in Figure 27 used data from the reference greenhouse, it is noteworthy that the model CM_{col} was less affected by the values of the thermal screen. This behaviour was consistent over the whole test dataset.

The presented test periods make clear that the models gave a very good estimation of both climate variables in those cases where the actuators remained without change. The degree to which these changes affected each particular model was different, though. In general, the model trained with data from the collector greenhouse reacted more abruptly to changes in the ventilation opening, while the other two models reacted more to changes in the thermal screen closure.

AQUIVOY

4.2 ANN FOR SIMULATION PHYTOMETRIC SIGNALS

4.2.1 *Model Design*

The model PM comprised three independent ANN, for which reason this section presents the important design results in parallel. Since these networks did not incorporate previous time steps as inputs, the design process was limited to the selection of network architecture (number of hidden nodes) and input signals.

In general, the errors achieved by all the submodels of PM were bigger than those shown by climate prediction. While the latter were in all cases in the range $(1 \times 10^{-4}, 4 \times 10^{-4})$ (Figure 15), the errors in the simulation of phytometric signals remained between 1×10^{-3} and 5×10^{-3}, i.e. one order of magnitude above. The errors for PM are shown in Figures 28 to 30.

There were other important differences between CM and PM. When simulating phytometric signals:

- The models trained with all available inputs always outperformed their counterparts only fed with the minimum set.
- The variability of each single model was much lower (in comparison with the magnitud of the errors). This led to the error bars (showing the 95% confidence interval of the mean) in Figures 28 to 30 to be partly hidden behind the markers.
- The MSE consistenly decreased with an increasing number of hidden neurons.

The networks with more input nodes ($n_h = 2 \cdot n_i + 1$) gave the least MSE and this architecture (ANN_3) was thus selected for the three submodels of PM. Nonetheless, the results of the variance analysis showed no significant differences between architectures.

The diagrams in Figures 31 to 34 show the relative effects of individual inputs on PMa, PMb and PMc. These diagrams are analog to those used for climate prediction (Figure 17).

The inputs that most improved the simulation of leaf temperature were the CO_2 concentration, the leaf area index and the outside air temperature. As shown in Figures 31 and 32, these three signals were consistently better in all three models (PMa_{col}, PMa_{ref} and PMa_{both}). Although other inputs peformed better ocasionally, it is worth nothing that this happend only in some of the models. Two examples of such inputs are the solar radiation, which had a strong effect in the model PM_{col} and the outside relative humidity, which had it in PM_{ref}.

The most important inputs for simulation of transpiration were the CO_2 concentration, the leaf area index and the leaf temperature (Figure 33). The effect of single inputs on the model performance was less

noticeable in the model PM_{col}, where every extra input included contributed to lower the simulation error in roughly the same amount. There is a strong contrast with the model PM_{ref}, where the effect of the named inputs (CO_2 concentration, the leaf area index and the leaf temperature) was very clear. The transpiration model was the only one with different inputs between greenhouses (see Table 12)

In the simulation of photosynthesis, the three inputs that mostly helped to improve the models were the CO_2 concentration, the leaf temperature and the transpiration rate (Figure 34). In all three models, the largest improvement was due to the inclusion of the transpiration rate. The leaf area index and the outside weather conditions showed a very moderate improvement in comparison with the three mentioned input variables and were hence dropped from the selected inputs.

Figure 28: PMa: Simulation of leaf temperature. MSE achieved after 500 training epochs, using different network architectures and datasets. The points indicate the mean of 15 MSE values (independent trainings). The 95% confidence interval of the mean is included as error bars on the points. See text for further details.

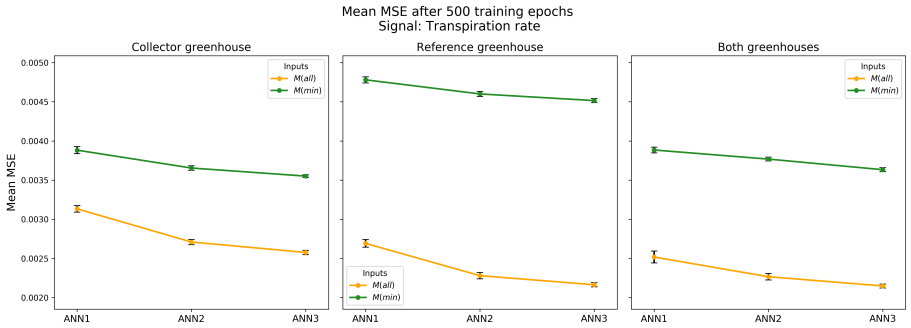


Figure 29: PMb: Simulation of transpiration rate. MSE achieved after 500 training epochs, using different network architectures and datasets. The points indicate the mean of 15 MSE values (independent trainings). The 95% confidence interval of the mean is included as error bars on the points. See text for further details.

Figure 30: PMc: Simulation of photosynthesis rate. MSE achieved after 500 training epochs, using different network architectures and datasets. The points indicate the mean of 15 MSE values (independent trainings). The 95% confidence interval of the mean is included as error bars on the points. See text for further details.

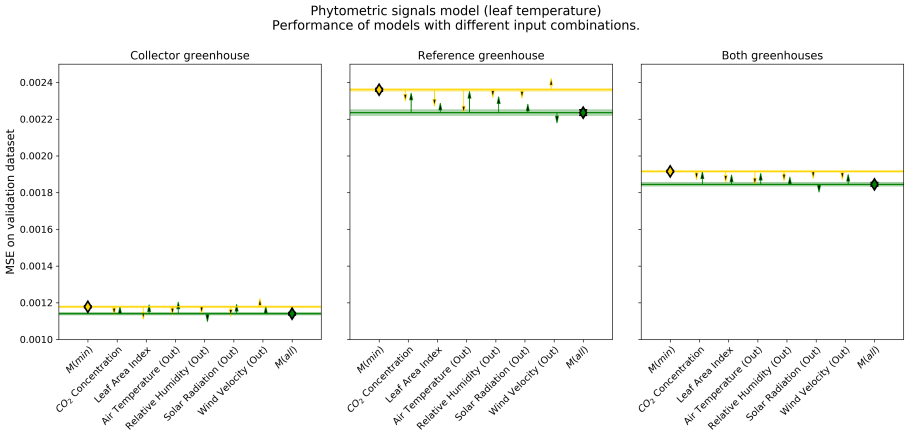


Figure 31: The selected network architecture for PMa (leaf temperature), trained with different combinations of inputs. The 95% confidence interval for the means (yellow for M_A and green for M_B) is very narrow and can only be appreciated as the coloured stripes around the horizontal lines, the error bars remaining hidden behind the diamond markers. The yellow arrows show the error of a model incorporating the input i. The green arrows show the error of a model lacking the input i.

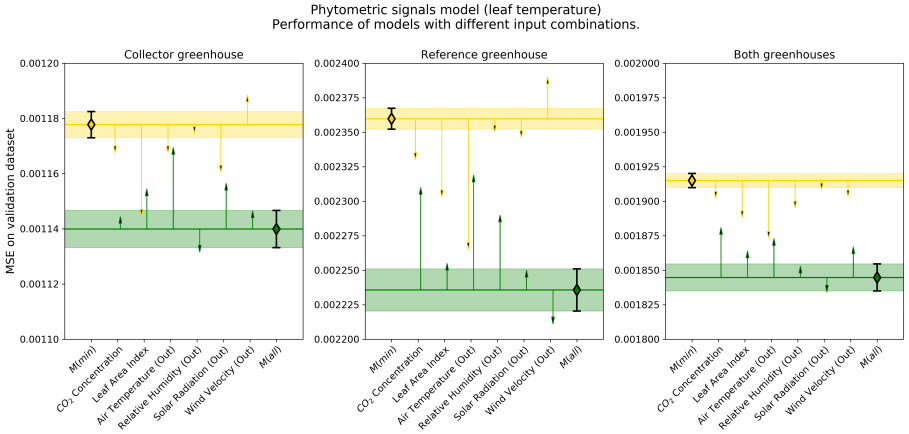


Figure 32: A close-up of the diagrams in Figure 31. Note the different scale of the y-axes.

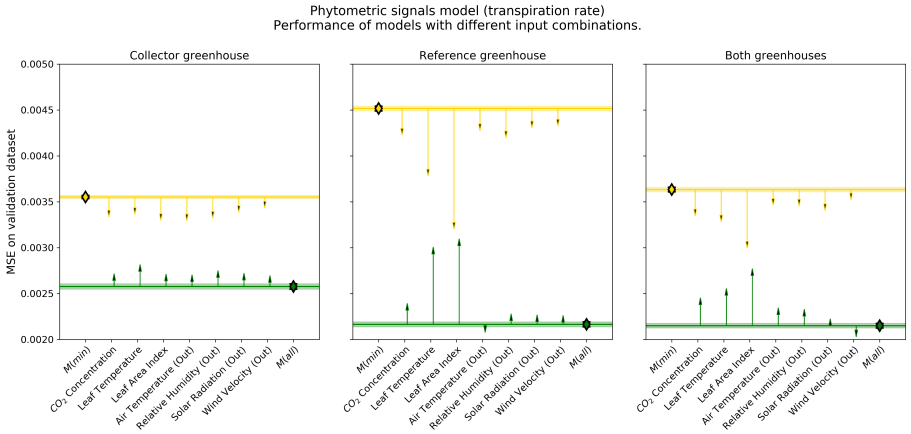


Figure 33: The selected network architecture for PMb (transpiration rate), trained with different combinations of inputs. The 95% confidence interval for the means (yellow for M_A and green for M_B) is very narrow and can only be appreciated as the coloured stripes around the horizontal lines, the error bars remaining hidden behind the diamond markers. The yellow arrows show the error of a model incorporating the input i. The green arrows show the error of a model lacking the input i.

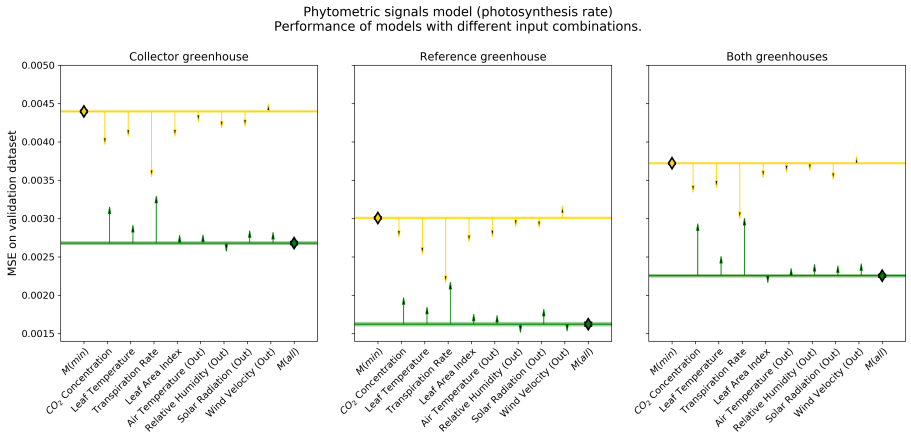


Figure 34: The selected network architecture for PMc (photosynthesis rate), trained with different combinations of inputs. The 95% confidence interval for the means (yellow for M_A and green for M_B) is very narrow and can only be appreciated as the coloured stripes around the horizontal lines, the error bars remaining hidden behind the diamond markers. The yellow arrows show the error of a model incorporating the input i. The green arrows show the error of a model lacking the input i.

The inputs selected for the three networks in the phytometric signals model are summarised in Tables 11 to 13. In all cases, the selected number of hidden nodes was $n_h = 2 \cdot n_i + 1$ (ANN_3).

	Phytometric signals Leaf Temperature Model		
Inputs selected	PMa_{col}	PMa_{ref}	PMa_{both}
Air temperature	✓	✓	✓
Relative humidity	✓	✓	✓
Hour angle	✓	✓	✓
Solar declination	✓	✓	✓
Solar elevation	✓	✓	✓
Theoretical solar radiation	✓	✓	✓
CO_2 concentration	✓	✓	✓
Leaf Area Index	✓	✓	✓
Outside air temperature	✓	✓	✓
Outside relative humidity	✗	✗	✗
Outside solar radiation	✗	✗	✗
Outside wind velocity	✗	✗	✗

Table 11: Inputs selected for the leaf temperature model.

The learning curves for the three networks PMa, PMb and PMc are shown in Figures 35 to 37. It took 10 to 20 training epochs for the error curves to stabilise their path towards an error minimum. As in CM, these learning curves do not show any signs of overfitting or other training problems.

Inputs selected	Phytometric signals Transpiration Rate Model		
	PMb_{col}	PMb_{ref}	PMb_{both}
Air temperature	✓	✓	✓
Relative humidity	✓	✓	✓
Hour angle	✓	✓	✓
Solar declination	✓	✓	✓
Solar elevation	✓	✓	✓
Theoretical solar radiation	✓	✓	✓
CO_2 concentration	✓	✓	✓
Leaf Area Index	✓	✓	✓
Outside air temperature	✓	✗	✗
Outside relative humidity	✓	✗	✗
Outside solar radiation	✓	✗	✗
Outside wind velocity	✗	✗	✗
Leaf temperature	✓	✓	✓

Table 12: Inputs selected for the transpiration rate model.

Inputs selected	Phytometric signals Photosynthesis Rate Model		
	PMc_{col}	PMc_{ref}	PMc_{both}
Air temperature	✓	✓	✓
Relative humidity	✓	✓	✓
Hour angle	✓	✓	✓
Solar declination	✓	✓	✓
Solar elevation	✓	✓	✓
Theoretical solar radiation	✓	✓	✓
CO_2 concentration	✓	✓	✓
Leaf Area Index	✗	✗	✗
Outside air temperature	✗	✗	✗
Outside relative humidity	✗	✗	✗
Outside solar radiation	✗	✗	✗
Outside wind velocity	✗	✗	✗
Leaf temperature	✓	✓	✓
Transpiration rate	✓	✓	✓

Table 13: Inputs selected for the photosynthesis rate model.

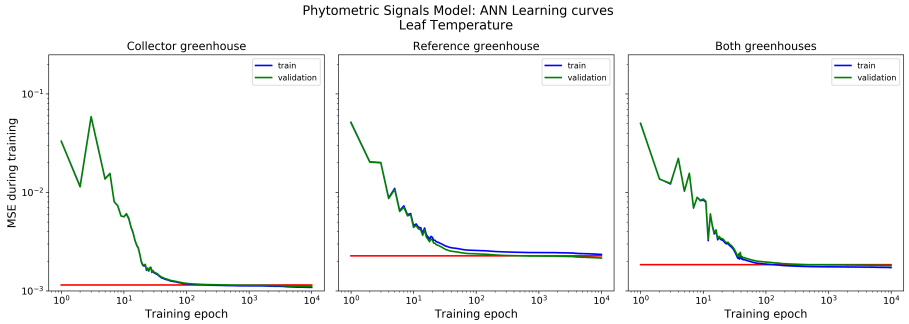


Figure 35: PMa: Leaf temperature. Training of the final models over 10,000 epochs. The red lines show the MSE at 500 training epochs, as a reference.

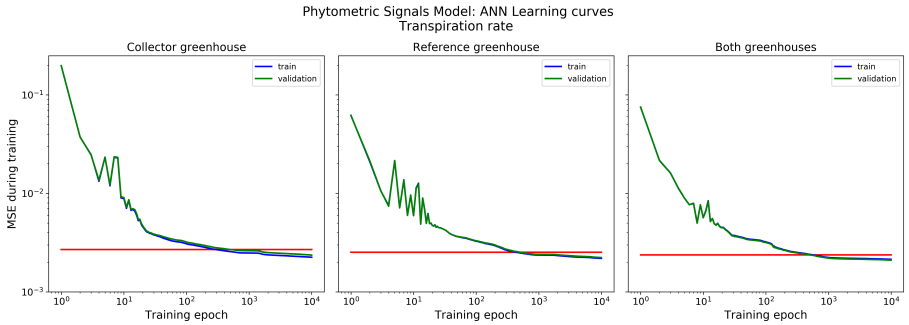


Figure 36: PMb: Transpiration rate. Training of the final models over 10,000 epochs. The red lines show the MSE at 500 training epochs, as a reference.

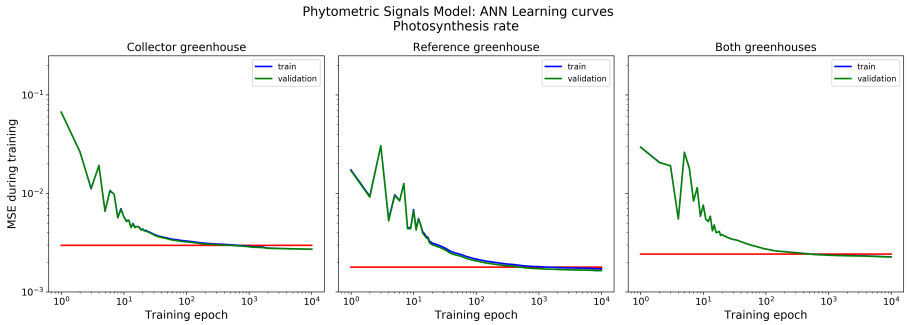


Figure 37: PMc: Photosynthesis rate. Training of the final models over 10,000 epochs. The red lines show the MSE at 500 training epochs, as a reference.

4.2.2 *Model Test*

Overall results

For the phytometric signals, the combined model (PM_{both}) was more robust across test greenhouses, giving nearly as good results as the corresponding models: For all effects and purposes, it simulates the collector greenhouse as good as the model PM_{col} and the reference greenhouse as good as the model PM_{ref}.

Table 14 gives a summary of the coefficients of determination (R^2) resulting from the test of PM. This table is included to serve as a quick reference, since the scatterplots that follow will give more detailed insights on the results from each of the models.

Like in CM, PMa, PMb and PMc trimmed the simulation results close to both ends of the scale (Section 4.1.2). This led to the models overestimating small values and underestimating big ones. While this happened in the simulation of the three phytometric signals, there were cases when the measurements did not reach the limits at all, for example: The photosynthesis measured in the reference greenhouse did not reach this upper limit at all.

The leaf temperature results, as presented in Figure 38, show the above-mentioned effect: The points deviated from the 45°-line, towards an *s-shaped* form. The deviation is more apparent in the upper end of the curve, but it can be seen in the lower end of the left panel as well. Note that leaf temperatures under $15^{\circ}C$ occurred rather scarcely in the reference greenhouse.

All leaf temperature models performed better in the collector greenhouse than in the reference greenhouse. Not only does the left panel in Figure 38 present a more compact, band-like, point cloud, but also the Figure 39 shows more compact box-and-whisker plots for the collector greenhouse.

The model PMa_{ref} was able to simulate the leaf temperature better in the collector than in the reference greenhouse. Conversely, all models tended to underestimate the leaf temperature in the reference greenhouse, when the measurements went beyond $25^{\circ}C$.

Variable	Test greenhouse	Model		
		PM_{col}	PM_{ref}	PM_{both}
Leaf temperature	Collector	0.979	0.958	0.978
	Reference	0.893	0.913	0.914
Transpiration rate	Collector	0.642	0.793	0.849
	Reference	0.588	0.927	0.927
Photosynthesis rate	Collector	0.821	0.788	0.810
	Reference	0.835	0.867	0.860

Table 14: Coefficients of determination (R^2) for all models in PM, as tested in two greenhouses.

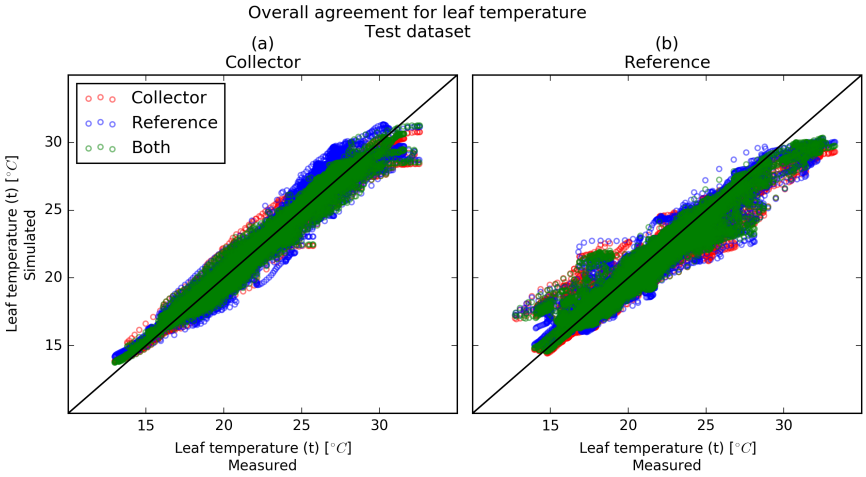


Figure 38: Measured and simulated leaf temperature. Left: Collector greenhouse. Right: Reference greenhouse. The colours indicate the dataset used for training.

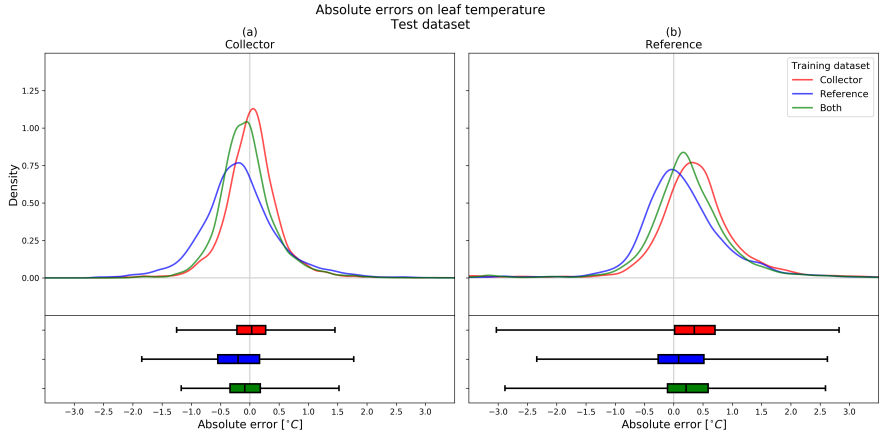


Figure 39: Absolute errors on the simluation of leaf temperature. Top: Kernel density estimation of the errors. Bottom: Distribution of the same errors. The whiskers mark the percentiles [1,99]. Left: Collector greenhouse. Right: Reference greenhouse. The colours indicate tthe dataset used for training.

In contrast to the leaf temperature, all transpiration models performed better in the reference than in the collector greenhouse. When tested in the collector greenhouse, the model PMb_{ref} was prone to overestimate the transpiration, while the PMb_{col} rather underestimated it. This behaviour can be seen in Figure 40 as a blue and a red point cloud, respectively. It can also be seen in Figure 41, in the form of asymmetrical whiskers in the boxplots.

As a matter of fact, the model PMb_{col} strongly underestimated the transpiration rate, regardless of the test greenhouse. However, the underestimation error had a lower incidence in the collector greenhouse, because the measured transpiration rate did not surpass the $120 mg \cdot s^{-1} \cdot m^{-2}$ there.

In the simulation of transpiration, the combined model PMb_{both} was able to reach a lower absolute error in both greenhouses. This is more clearly seen in both panels of Figure 41, which depict a lower and more symmetric absolute error for this model.

Like the transpiration, the photosynthesis rate was simulated better in the reference greenhouse. In this case, there was little difference between the three models.

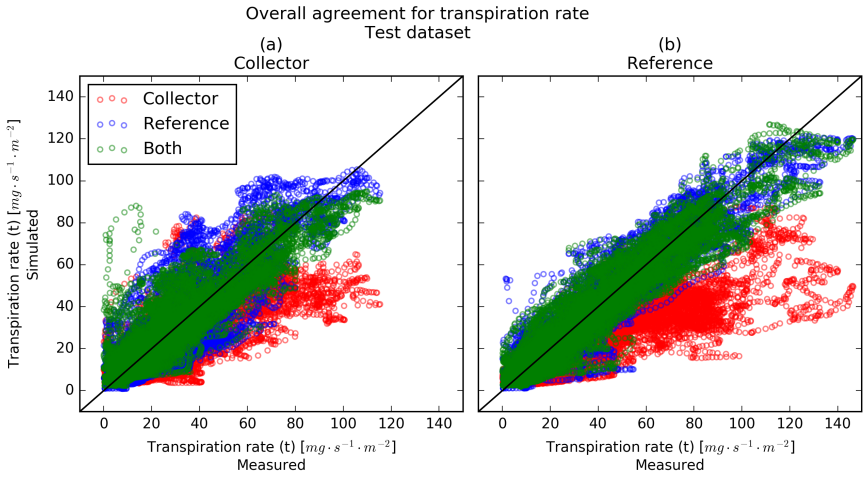


Figure 40: Measured and simulated transpiration rate. Left: Collector greenhouse. Right: Reference greenhouse. The colours indicate the dataset used for training.

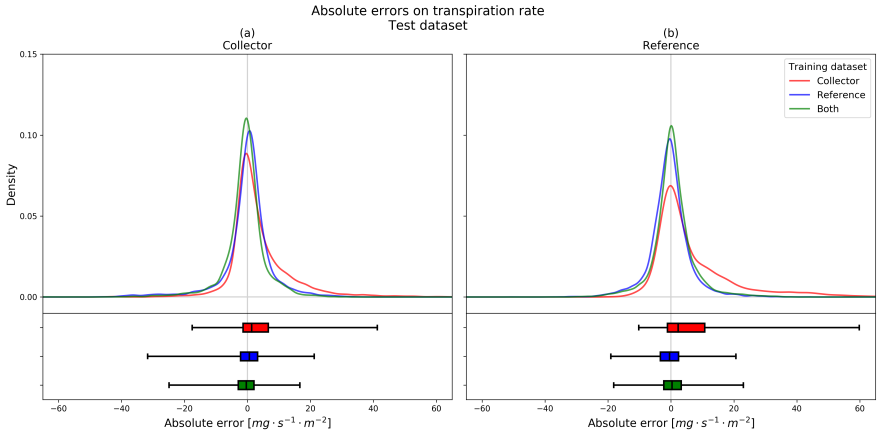


Figure 41: Absolute errors on the simluation of transpiration rate. Top: Kernel density estimation of the errors. Bottom: Distribution of the same errors. The whiskers mark the percentiles [1,99]. Left: Collector greenhouse. Right: Reference greenhouse. The colours indicate the dataset used for training.

The presence of high values of photosynthesis in the collector greenhouse was reflected in a cloud of underestimated values, by all three models. Also, the model PMc_{ref} overestimated several values between 5 and $15\mu g \cdot s^{-} \cdot m^{-2}$.

It is also noteworthy that the measurements above $20\mu g \cdot s^{-1} \cdot m^{-2}$ were rather scarce in the reference greenhouse. The measured values thus did not reach the point at which the model restricts the simulated results. By contrast, this value can be seen in the collector greenhouse as the horizontal limit around that value. This phenomenon was already shown in the transpiration rate (Figure 40), whereas the high transpiration values presented themselves in the reference greenhouse instead.

The results achieved by the combined models (PM_{both}) were overall better for the three phytometric signals. In general, the single-greenhouse models performed slightly better when tested in the corresponding greenhouse, that is, the PM_{col} models made better simulations in the collector than in the reference greenhouse, and vice versa. Two notable exceptions were:

- **Leaf temperature:** The model trained with data from the reference greenhouse (PMa_{ref}) yielded better simulations in the collector than in the reference greenhoouse.
- **Transpiration rate**: The model trained with data from the collector greenhouse (PMb_{col}) performed better in the reference greenhouse. Additionally, the models PMb_{ref} and PMb_{both} simulated the collector greenhouse even better than PMb_{col}.

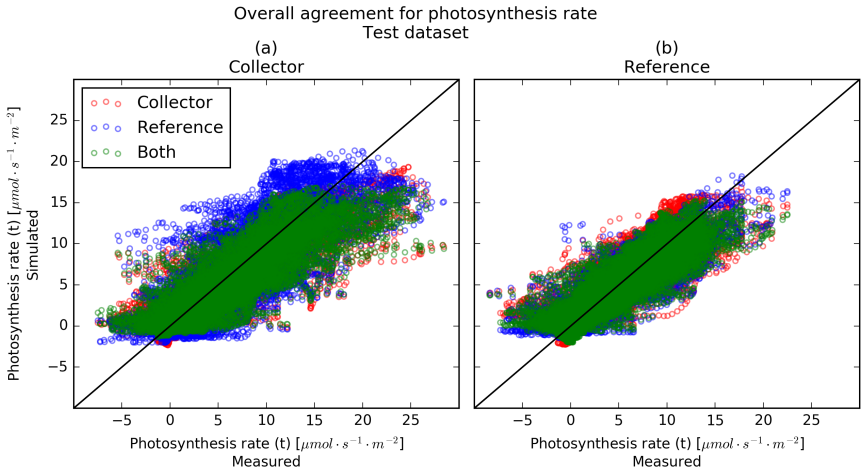


Figure 42: Measured and simulated photosynthesis rate. Left: Collector greenhouse. Right: Reference greenhouse. The colours indicate the dataset used for training.

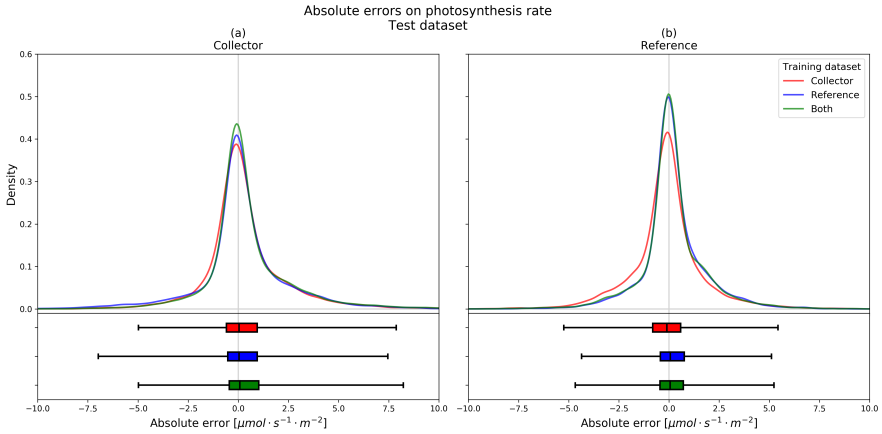


Figure 43: Absolute errors on the simulation of photosynthesis rate. Top: Kernel density estimation of the errors. Bottom: Distribution of the same errors. The whiskers mark the percentiles [1,99]. Left: Collector greenhouse. Right: Reference greenhouse. The colours indicate the dataset used for training.

Selected Test Periods

The following figures present a selection of simulations, more specifically 4 series of simulations, each consisting on 2 consecutive days of data. The 2 first test periods (Figures 44 and 45) are included because they show the general behaviour of the simulations. The next 2 test periods (Figures 47 and 48) make use of an exceptional situation (a period when the signals were out of service during several days) to show the simulation of transpiration and photosynthesis using already simulated inputs.

The measured and simulated signals in Figure 44 help to visualize several patterns already mentioned in the previous section:

- The leaf temperature was underestimated at high values.
- The model PMb_{col} tended to underestimate the transpiration rate.
- Very high values of photosynthesis (most notably in the collector greenhouse) tended to be underestimated.

The label **A** in the panels 1 and 2 of Figure 44 indicates high leaf temperature values that were underestimated by the models. The values on the far-right end of the scatterplots in Figure 38 belong to this class of error. Apart from these cases, the leaf temperature simulations gave very good results.

The panels 3 and 4 of the same figure show simulations where the model PMb_{col} underestimated the transpiration rate, a fact that was already mentioned in previous sections (compare with Figures 40 and 41). With very few exceptions, the simulations made with the model PMb_{both} lay in magnitude between those by the other two models. As stated, the PMb_{col} model was prone to underestimate, whereas the model PMb_{ref} overestimated the transpiration more often.

The peak marked with **B** in Figure 44 is one of a series of measurements that are conditioned by rapid changes in the relative humidity. In fact the oscillations in the transpiration during the 17th of May (collector greenhouse) were closely related with the first and second derivative of the relative humidity. While the simulated signals also showed this pattern, their oscillations were less intense.

On the other hand, the causes of the valley tagged as **C** could not be identified. A close analysis of the data did not reveal any particular changes in actuators or input variables, nor in the weather conditions at that particular point. A possible cause could be a mechanical disconnection or rearrangement of the leaf cuvettes. Note that the models PMb_{ref} and PMb_{both} seemed to bypass this period, carrying on with the upwards morning transpiration trend.

The photosynthesis rate measurements were characterised by stong oscillations, especially in the collector greenhouse. These oscillations (marked in both greenhouses as **D**) were closely related with rapid changes in the concentration of CO_2 (they correlated with the first and second derivatives of these values) due to the measuring technique. Since the models did not take into account the previous values of the signal, the simulated signals apparently overlooked these oscillations, resulting in smoother curves, but also in series of under- and overestimated values. These situations, when the simulations fell between high and low measurements, also gave form to the scatterplot presented in the previous section (Figure 42).

The run charts in Figure 45 give a second view on the phytometric signals simulations. The two first panels in this figure, show how the leaf temperature simulations closely follow this singal's measurements. In this test period (as opposed to that in Figure 44), no major underestimation occurred. The transpiration rate in both greenhouses was underestimated by the model PMb_{col}, particularly on the 4th of August. The photosynthesis, too, presented a pattern similar to that in Figure 44: Strong oscillations of the measurements in the collector greenhouse in the daytime, as well as peaks in the early morning in the reference greenhouse. The photosynthesis simulations followed smoother paths, especially in the reference greenhouse. In the collector greenhouse, this smoothing led to all three models "cutting through" the oscillations, thus underestimating the highest measured values.

However, the most remarkable trait in Figure 45 is a period (marked as **X**) when the measuring system was out of service in the collector greenhouse. This interruption of the sensor measurements affected the three phytometric signals overnight, from 21:20 to 05:00 the next morning. The lack of measurements is more noticeable in the panel 1 (leaf temperature) because the reported measurement was further away from

the expected values. The panels 3 and 5 show the transpiration and photosynthesis rates, simulated with the erroneous leaf temperature measurement as input (in the collector greenhouse), hence leading to the drift in photosynthesis (panel 5) during this period.

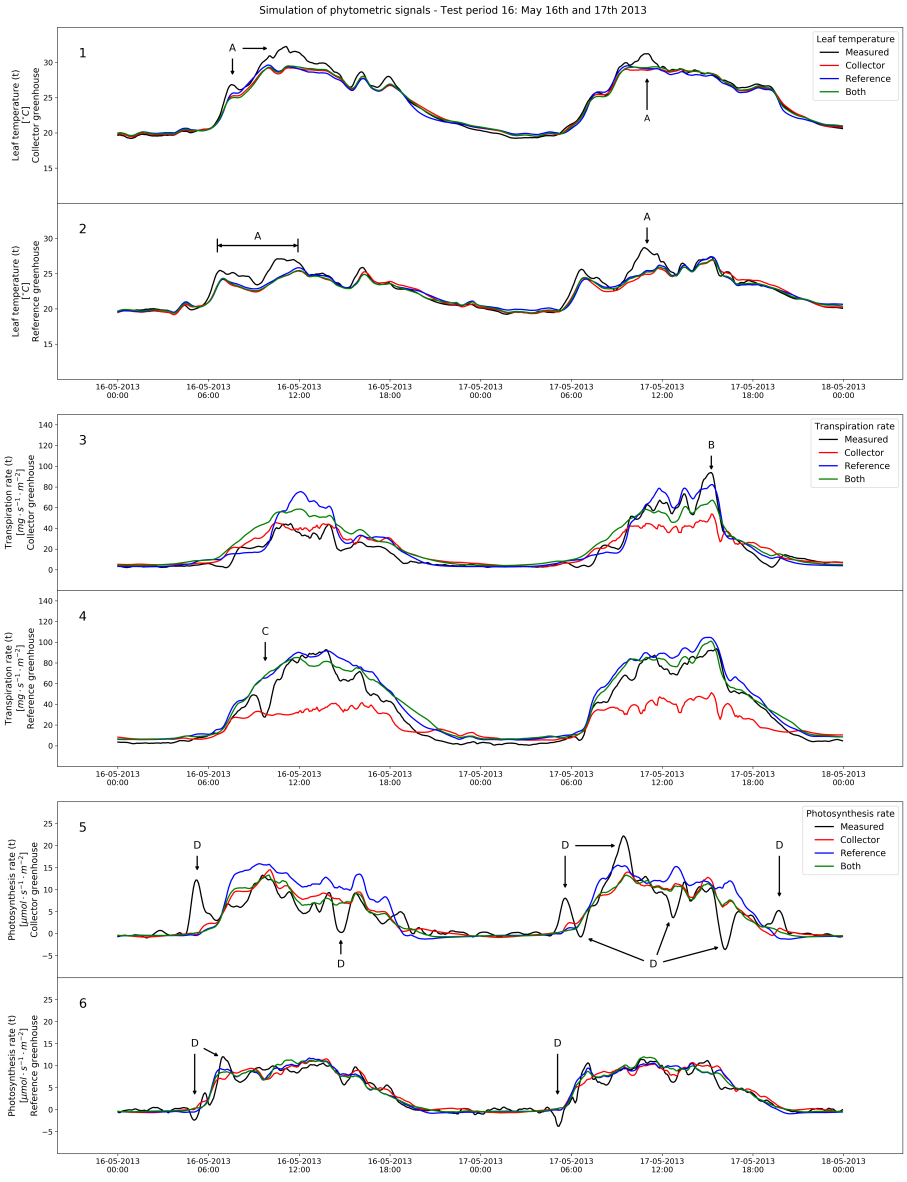


Figure 44: Simulation of phytometric signals: June 16th and 17th 2013. 1, 2: Leaf temperature. 3, 4: Transpiration rate. 5, 6: Photosynthesis rate. The coloured lines refer to the dataset used for training.

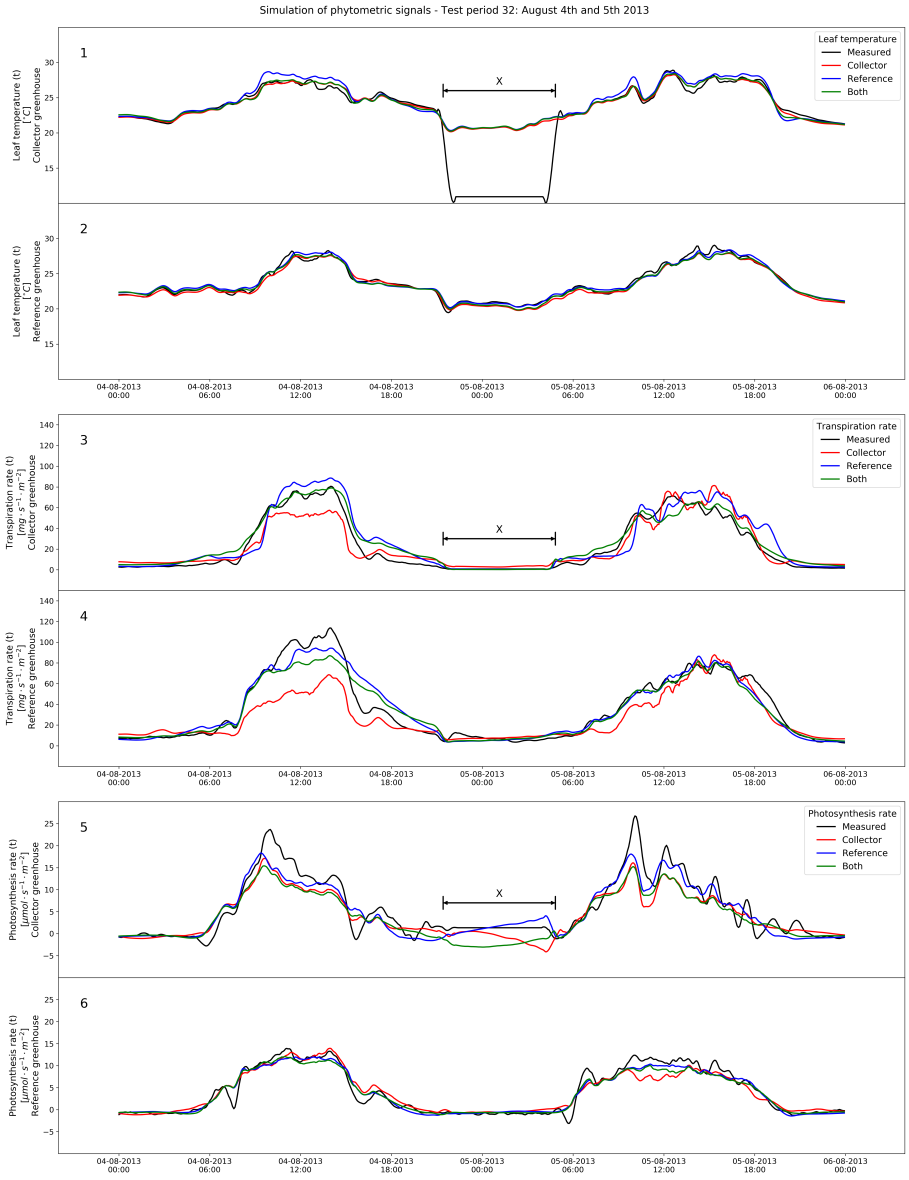


Figure 45: Simulation of phytometric signals: August 4th and 5th 2013. 1, 2: Leaf temperature. 3, 4: Transpiration rate. 5, 6: Photosynthesis rate. The coloured lines refer to the dataset used for training.

Simulating Using Simulated Inputs

The next 2 figures deal with a case study of particular interest. They show an interruption of the sensor measurements, similar to that in Figure 45, but with the following peculiarities:

- Only the reference greenhouse was affected
- The signals were absent during day and night
- The measurements at the plants were affected at different time periods:
  - All phytometric measurements fell out simultaneously (on the early morning of June 6th 2013)
  - Other signals were not affected (climate, actuators, weather)
  - The leaf temperature measurements were the first to be restored (on the morning of June 10th 2013)
  - The measurements of transpiration and photosynthesis remained absent until June 13th

The Figures 47 and 48 mark the period with no phytometric measurements with the label **X**. This annotation line spans over the absence of the leaf temperature only, since the transpiration and photosynthesis measurements were not present for the rest of the depicted time. In other words, the three signals were working correctly, then all three failed, the leaf temperature was thereafter restored, leaving the transpiration and photosynthesis out of service until the end of the depicted time.

In these examples, only simulations from model PM_{both} are shown. This model was used to provide two signals:

1. With measured input variables, which implies the use of erroneous values of leaf temperature and transpiration rate as inputs.
2. With simulated values of leaf temperature and transpiration rate as model inputs.

Since the leaf temperature did not rely on previous phytometric signal simulations, the corresponding top panels only show one simulated signal.

The cascade configuration used to obtain simulations without actual measurements (which was already shown on Figure 13 in its generic form) is shown again in Figure 46, with regard to the specific input signals selected for the model PM_{both} (Tables 11 to 13).

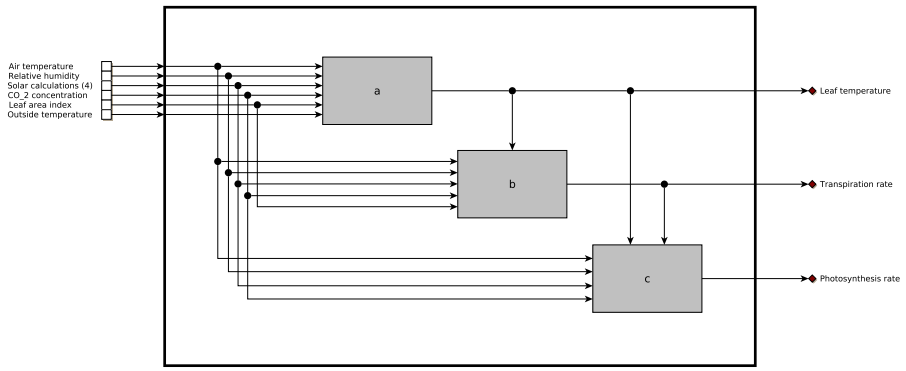


Figure 46: Phytometric signal models PM_{both} connected in cascade.

The points at which the leaf temperature measurements fell out and started working again mark differences in the simulation of transpiration and photosynthesis. Whenever the measurements of leaf temperature were reliable, they enhanced the simulations of the other two variables.

On the other hand, when no reliable measurements of leaf temperature were available, the simulations of transpiration and photosynthesis were strongly affected (period labeled **X**) by the erroneous values. In these cases the simulated leaf temperature provided a means to carry on the simulations.

Above all, the two simulation modes diverged in the photosynthesis signals, as seen in the periods before and after the **X** label, as well as in the panels for the collector greenhouse. The differences in the simulations of photosynthesis were more pronounced than those of transpiration, partly because the simulated transpiration already carried an error forward to the inputs of the photosynthesis model.

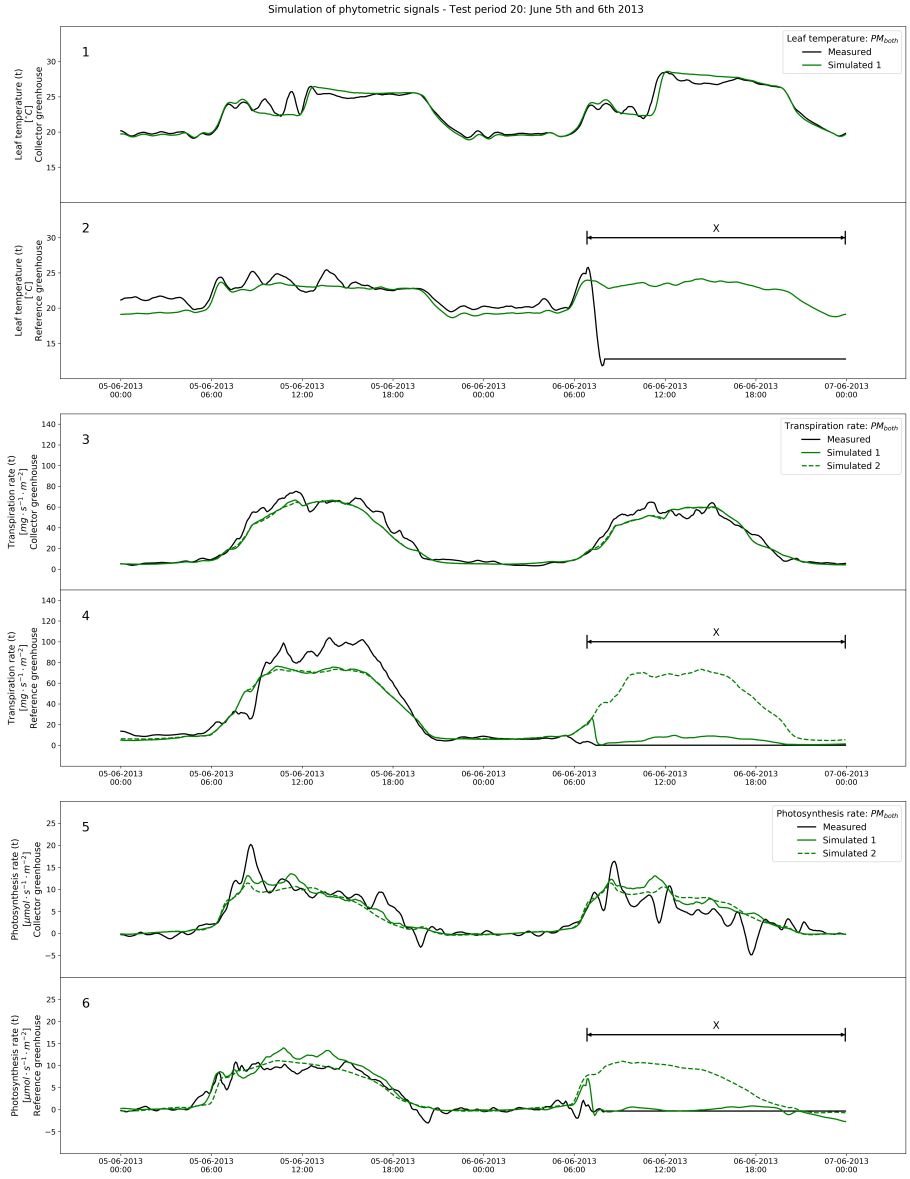


Figure 47: Cascade simulation of phytometric signals: June 5th and 6th 2013. 1, 2: Leaf temperature. 3, 4: Transpiration rate. 5, 6: Photosynthesis rate. The simulations marekd as *Simulated 1* in the panels 3 to 6 were made with phytometric signals as inputs, while those marked as *Simulated 2* took simulated values as inputs.

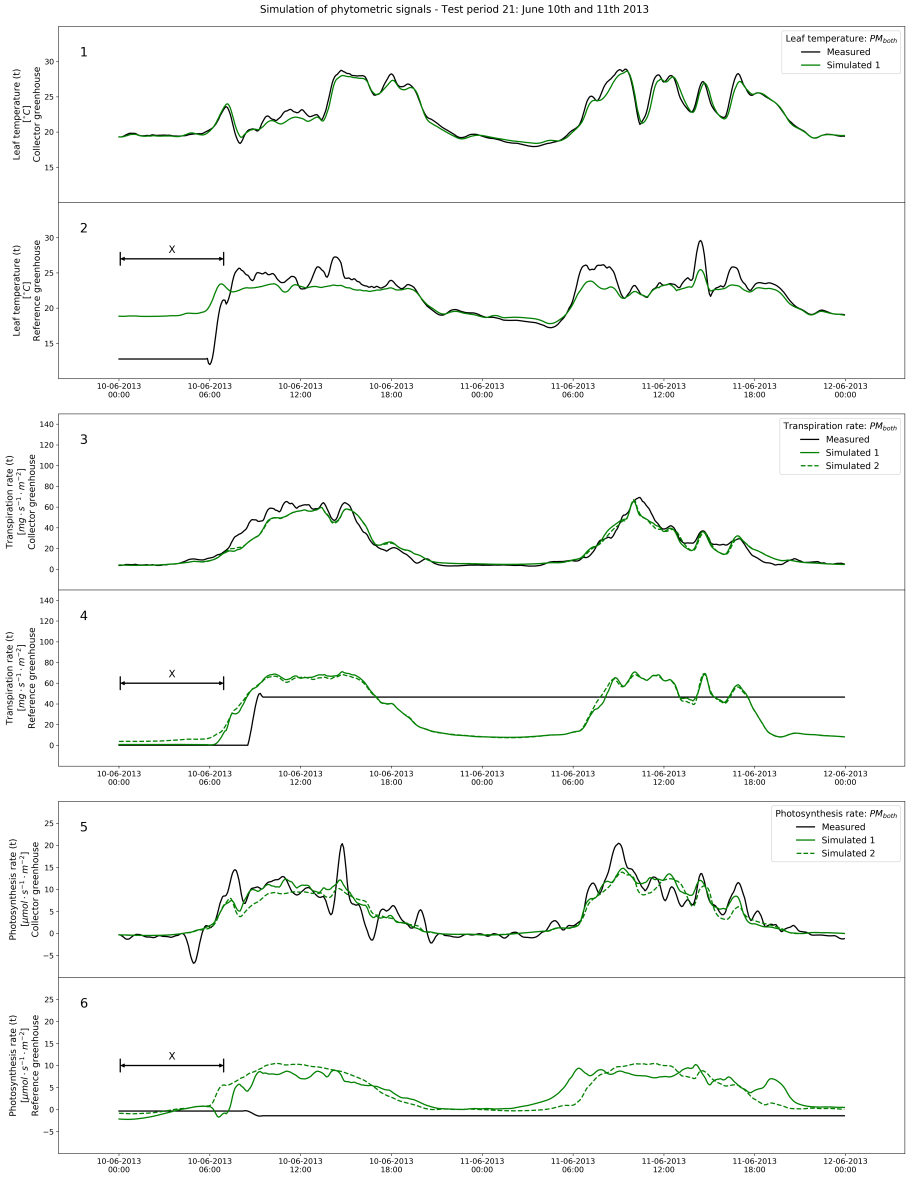


Figure 48: Simulation of phytometric signals: June 10th and 11th 2013. 1, 2: Leaf temperature. 3, 4: Transpiration rate. 5, 6: Photosynthesis rate. The simulations marekd as *Simulated 1* in the panels 3 to 6 were made with phytometric measurements as inputs, while those marked as *Simulated 2* took simulated values as inputs.

4.3 LONG-TERM PREDICTION BY MEANS OF RECURSION

4.3.1 *Model Coupling*

This section describes additional experiments carried on with the two independent models (climate prediction and phytometric signals) presented in Sections 4.1 and 4.2.

The experiments take advantage of the input-output configuration of the models, which allowed them to be coupled together to generate more complex simulations.

Two model couplings were tested:

- Using the output of the climate prediction model as input to itself: $CM \rightarrow CM$;
- Using the output of the climate prediction model as input to the phytometric signals model: $CM \rightarrow PM$

In the first case, the output from CM, that is, the one-step prediction (OSP) of climate, was fed back to the system; when done recursively this produced a long-term prediction (LTP) of the same signals. The block labelled **I** in Figure 49 represents the climate LTP.

When the simulated values of temperature and relative humidity were fed recursively to the model, the values of the actuators were actual records from the automation system. For each iteration, the solar coordinates were fed according to their respective theoretical values. Each LTP was limited to 6 prediction steps, representing a forecasting of 30 minutes.

The model PM, as described in Section 4.2, took the measured greenhouse air conditions as inputs to estimate the plant response. In Figure 49 this is represented by the block labelled **II**. Since PM does not change the estimation time, the results in **II** can be thought of as being online estimations of the measurements taken at the plants.

The phytometric signal estimations in the block **III** refer to future points in time. Once a climate LTP was generated, the output of each climate prediction could be used to estimate the plant response at the

corresponding time, thus generating a long-term prediction of the phytometric signals.

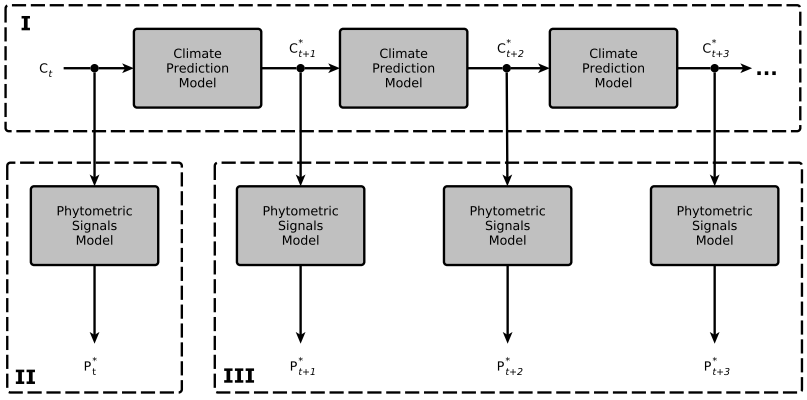


Figure 49: Using recursion to generate long-term predictions. I: Long-term prediction of climate. II: Simulation of phytometric signals using climate measurements. III: Long-term prediction of phytometric signals. All signals marked with $*$ are simulation results.

4.3.2 *Long-Term Prediction of Climate and Phytometric Signals*

This section presents results on the long-term prediction of climate inside a greenhouse as well as the corresponding plant responses to it. These simulations were made using only the models trained with data from both greenhouses: CM_{both}, PM_{both}. In the case of PM, the rates of transpiration and photosynthesis were simulated using already simulated inputs, as previously shown in Section 4.2.2. The complete test dataset was used, with exception of the last 30 minutes of each test day: Since the simulations predict up to 6 steps in the future, and a single time step represents 5 minutes, the last half hour of each day was dismissed as triggering point for a new simulation.

The results are presented as follows: First, a single simulation run, i.e. a single long-term prediction, is shown in Figure 50. Then a series of simulations run over a whole day follow, presented in Figures 51 and 52. Rounding off, Figure 53 gives an overall view of the simulation error over the predicted steps, regarding the complete test dataset.

To illustrate the LTP-simulation process, a single simulation can be seen in Figure 50. The arrows and squared markers in the top panel indicate the 6-steps climate LTP. This particular simulation was started with the measurements taken at 8:50, in the morning of April 17th, 2013 (as well as the two previous states at 8:40 and 8:45). The solid lines represent the data measurements available to run the simulation, while the dashed lines indicate the measurements taken afterwards.

The temperature predictions shown on top of Figure 50 followed a rather straight path, which made them deviate from the actual measurements as the number of predicted steps increased. By contrast, the relative humidity predictions started to deviate from the measurements, but lined up again with them after 6 simulation steps.

Tangential prediction paths were found in cases where the slope of the simulated variable changed its value rapidily, regardless of the actuator operation. The simulated air temperature in Figure 50 is an example of such a deviation.

Each climate simulation was used to estimate the corresponding phytometric signals, as shown in the lower panels of Figure 50. The measurements taken directly at the plants are shown as dashed lines. These measurements were not present in any form in the inputs to PM, being used only to calculate the prediction error.

In Figure 50 it is also shown that the leaf temperature closely follows the air temperature, both in the simulations and the measurements. In other words, the measured leaf temperature follows a path that is similar to that of the measured air temperature, and the simulated leaf temperature follows a path similar to that from the simulated air temperature. Similarly, the transpiration rate is strongly dependent on the relative humidity, although this dependency is clearer to see in Figure 52.

In many cases, the generation of LTP led to an increased error at each iteration step. This was an expected result, and is presented in Figure 51, where a series of independent LTP simulations are displayed one after the other in a one day timeline. A simulation was triggered at every time step, i.e. every 5 minutes, for a total of 282 simulations for each day-period in the test dataset. The ventilation opening and the thermal screen closure are included in Figure 51 to help exploring the errors and deviations in the simulations.

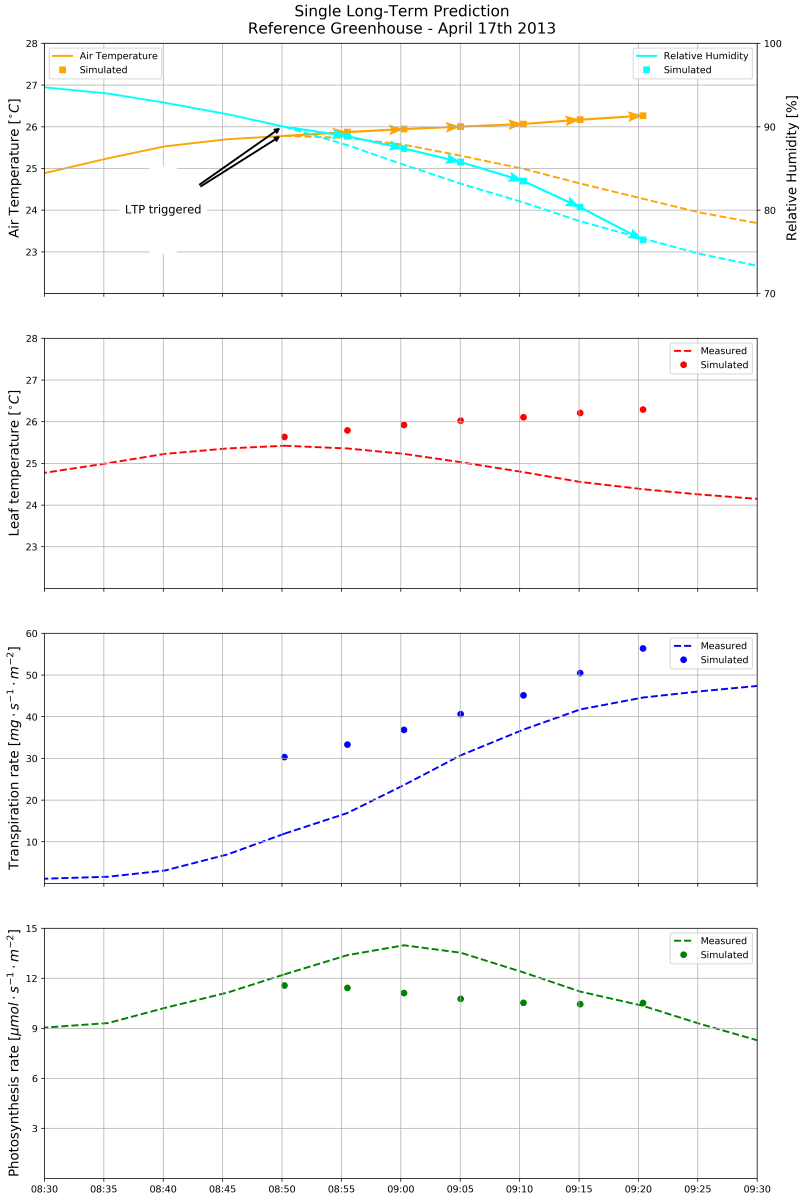


Figure 50: A single Long-Term Prediction, triggered at 8:50 with values from the reference greenhouse. The top panel shows the climate-LTP. The dashed lines indicate measurements not used as inputs for the simulations. The circles in the panels 2 to 4 are the simulated phytometric signals, which used the signals from the top panel.

The tangential deviation paths mentioned previously on page 108 are marked in Figure 51 with the tag **A**. These deviations occurred mainly with fluctuations of the simulated variable, also in cases when the actuators did not change their value.

The aforementioned deviations were much more apparent in the simulation of relative humidity. In particular, on the test period shown, the relative humidity measured on the reference greenhouse dropped rapidily in the morning, and continued to fluctuate during daytime. These fluctuations reflected themselves clearly on the LTP. Although these deviations were more striking in the reference greenhouse, the collector greenhouse was also affected, despite the fact that the relative humidity remained a more constant value over the day.

A second kind of error in the LTP simulations, marked in Figure 51 with the tag **B**, could be traced back to quick changes in the state of the actuators. This effect of the actuators over the simulations was already presented in previous sections (Section 4.1.2, particularly Figure 27). In these cases, a big inaccuracy in the first simulation (OSP) can be further propagated to the simulations that follow, thus leading to the mentioned error paths.

The phytometric signals simulations in Figure 52 build upon the climate LTP previously shown. The figure features predictions of leaf temperature, transpiration rate and photosynthesis rate for both greenhouses separately. In each panel, the corresponding LTP simulations of air temperature and relative humidity are included as reference.

In both greenhouses, the simulated leaf temperature followed the air temperature very closely. This close link between both signals implied that the simulation errors in the air temperature were consequently reflected in the simulated leaf temperature. This was particularly clear for those errors traced back to the fast operation of the thermal screen.

The simulated transpiration rate was strongly influenced by the relative humidity, although mirrored in trend: With increasing simulated values of relative humidity, the simulated values of transpiration sank. As observed with the leaf temperature, the errors in the simulated climate were also reflected in the simulated transpiration rate. The propagation of simulation errors from the climate to the transpiration is especially clear to see in the simulations from the reference greenhouse, because

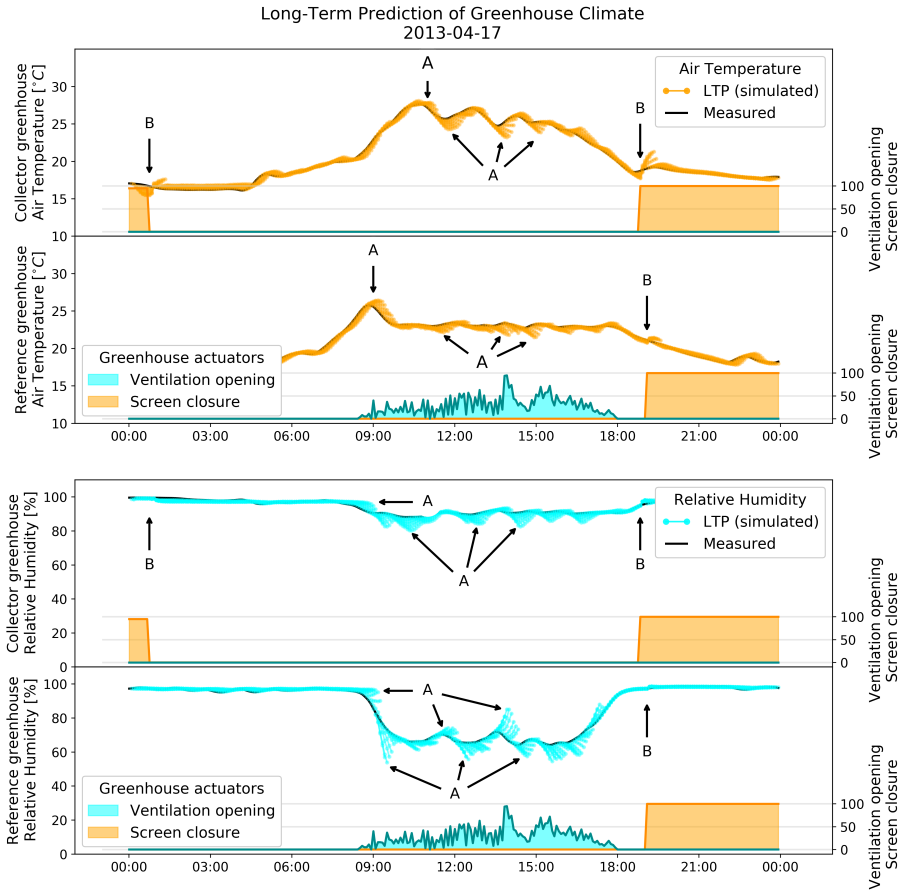


Figure 51: A series of 6-steps-LTP simulations of the greenhouse climate, triggered every 5 minutes. The right-hand axis shows the state of two greenhouse actuators affecting the simulations.

of the distinct error paths in the relative humidity during daytime. By contrast, the simulation errors due to the operation of the thermal screen did not seem to affect the transpiration noticeably, even though the errors were present in two of the the model inputs: Air temperature and leaf temperature.

A similar situation occurred with the photosynthesis: It did not show the big errors present in the inputs. This was remarkable, since its calculation used 4 already simulated inputs: Air temperature, relative humidity, leaf temperature and transpiration rates, all of which had strong oscillations at different points. Instead of reflecting errors present in the inputs, the simulated photosynthesis showed a rather smooth curve over the day.

Even though the photosyntesis model PMc_{both} appeared to automatically filter out many of the big errors present in its inputs, it yielded the lowest overall performance of all LTP simulations. In general, the errors in the photosynthesis LTP were less localised and could not be traced back to singular causes. The apparent filtering effect shown by the photosyntesis model had the side effect of missing fluctuations present in the measurements, therefore causing both positive and negative absolute errors (no clear over- nor underestimation were found).

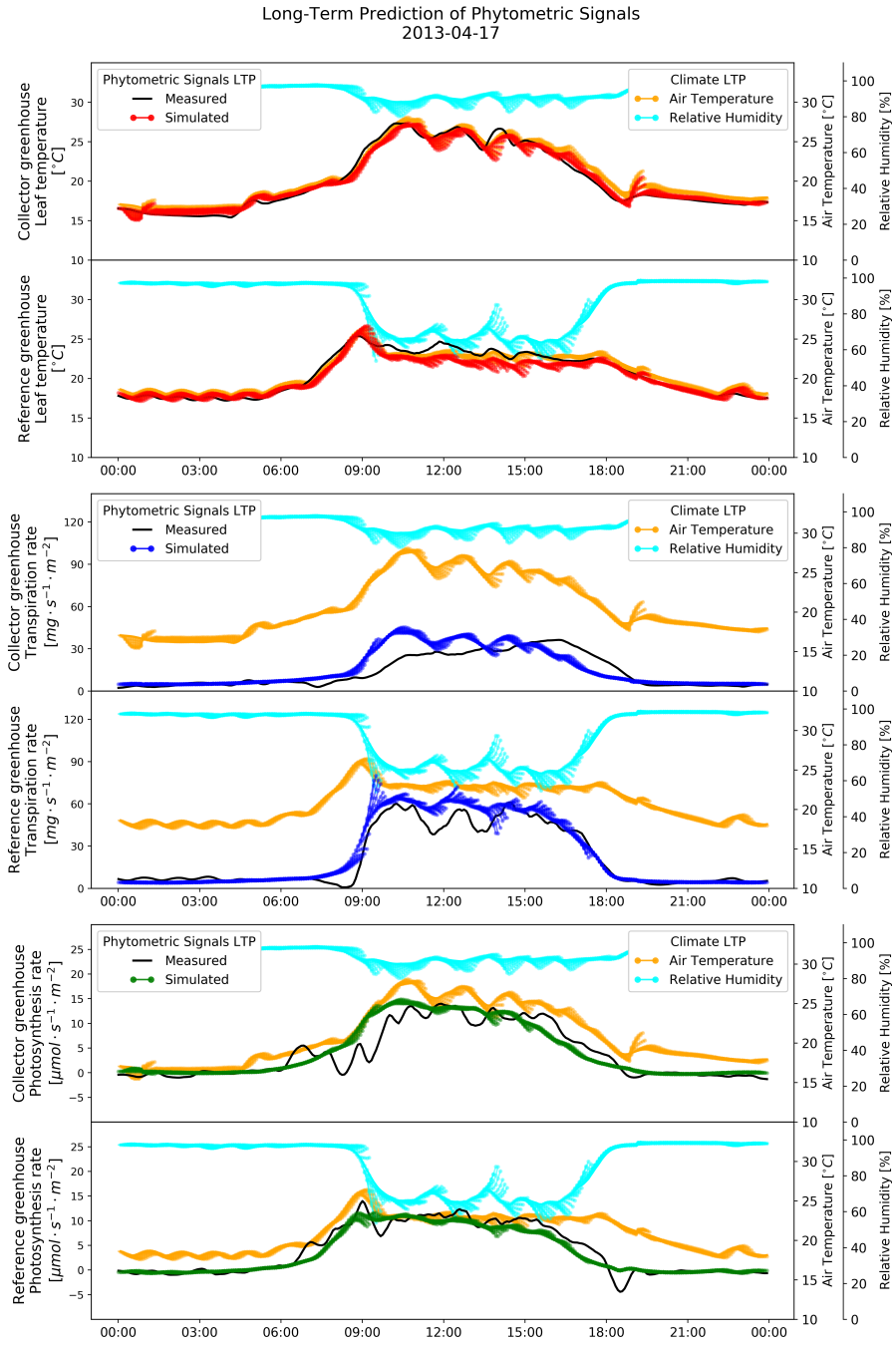


Figure 52: A series of 6-steps-LTP simulations of phytometric signals, triggered every 5 minutes. The climate LTP are included in all three panels as comparison.

A general view of the LTP error can be seen in Figure 52, in the form of three standard error measures: The coefficient of determination R^2, the root mean squared error $rmse$ and the mean absolute error (MAE). These measures of error were calculated using the whole test dataset, with exception of the last 6 records of each day period.

In all cases, the error increased monotonically with the number of simulated steps. The increasing error was of non-linear nature and was reflected consistently by the both error measures shown, although, this non-linearity was more conspicuous in the simulations of phytometric signals than in the simulations of climate.

Both climate variables, air temperature and relative humidity, showed a similar overall performance between greenhouses. The predictions of air temperature deteriorated slightly faster in the collector greenhouse, reaching an $rmse$ of $1.2^{\circ}C$ after 6 simuation steps, whereas the simulations from the reference greenhouse remained at $1.05^{\circ}C$. In contrast, the simulation of relative humidity yielded a very similar behaviour in both greenhouses.

Regarding the three phytometric signals, not only did the simulations differ more between greenhouses (as compared with the climate simulations), but also the non-linearity increase in the error was more apparent. Another difference between the climate prediction and the simulated phytometric signals was that in the former the error started to increase rapidily since the first simulation step, while in the latter it remained rather constant during the first simulation steps. In case of the transpiration rate, a clear increase in error could only be found after the third simulation ($t = +2$).

From the three simulated plant processes, the leaf temperature deteriorated faster with the number of prediction steps. Also, the combined model performed better when simulating the collector than the reference grenhouse, which contrasts with the transpiration and photosynthesis rates showing a smalller error in the reference greenhouse.

On the other hand, the photosynthesis simulations deteriorated slower: The $rmse$ between $t = 0$ and $t = +6$ increased 4.6% and 13.2% in the collector and reference greenhouse, respectively. In other words, the simulations of photosynthesis remained more stable throughout the prediction steps in both greenhouses than the other variables. The fact

that the error in photosynthesis increased slowly with the simulated steps can be seen as another expression of the smooth simulation paths shown in Figure 52 on page 113.

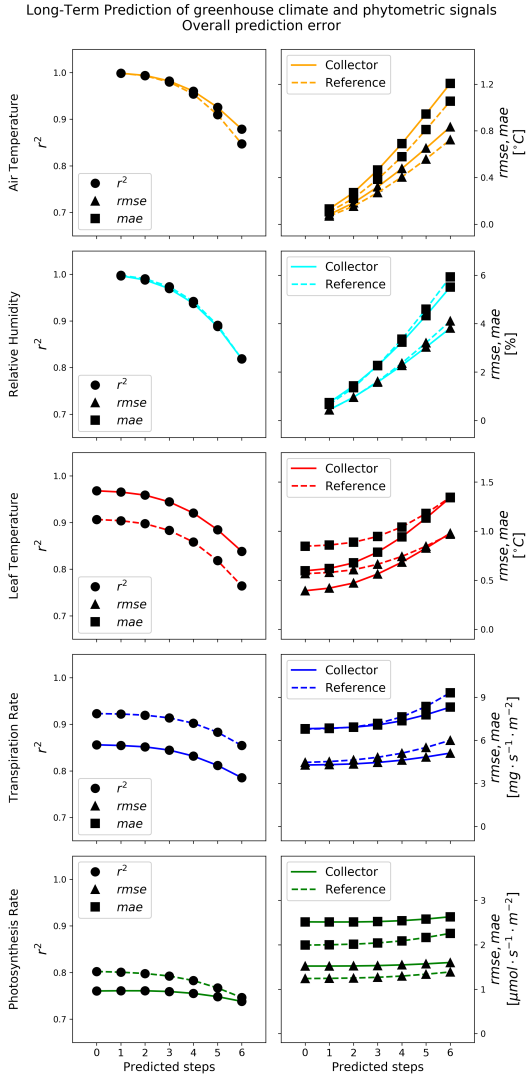


Figure 53: Two error measures over the complete test dataset against the number of predicted steps. Left: Coefficient of determination R^2. Right: Root mean square error ($rmse$) and Mean absolute error (MAE). The step 0, abscent in the climate simulations, is included in the phytometric signal panels to facilitate comparison.

Variable	Test greenhouse	
	Collector	Reference
Air temperature	821.61%	900.37%
Relative humidity	641.68%	795.73%
Leaf temperature	125.19%	58.74%
Transpiration rate	22.01%	37.37%
Photosynthesis rate	4.57%	13.24%

Table 15: Percentage of error increase ($rmse$) after a 6-steps LTP. For air temperature and relative humidity, the values refer to the 1st and 6th predicted step. For the three phytometric signals, they refer to the steps 0 and 6.

5

DISCUSSION

The aim of this dissertation was to implement and test two mathematical models of interest in greenhouse horticulture by means of ANN. The first model dealt with short-term prediction of the climate inside the greenhouse, and the second with the physiological plant reactions to their climatic environment. Due to the nature of the ANN, the focus of the calculations was set on the data available, both from direct measurements and model calculations, rather than causal or deterministic processes. Consequently, the design and test was oriented to the combination and subsequent comparison of datasets, but also to the structure and connectivity between the models.

This chapter discusses the relevance and context of the results presented previously, and is structured in accordance with Chapter 4:

- Section 5.1 discusses the climate prediction models, with a special focus on model design and the role of the greenhouse actuators
- Section 5.2 considers the results from the simulation of phytometric signals, with a particular interest in their potential practical applications
- Section 5.3 reviews the long-term prediction of climate and phytometric signals, as an example of model connectivity and stressing the opportunities for new research directions

5.1 GREENHOUSE CLIMATE PREDICTION

The model design presented in Section 4.1 started with an assessment of the inherent variability of the outputs from the ANN, which is a consequence of the random initialisation of the network weights described in Section 2.2.1 on page 28. The fact that a model with a given architecture can be optimally calibrated in different ways, i.e. the optimisation

algorithm can lead to different local minima (of the simulation error), is of importance for model selection and comparison. Even though there is a number of models based in ANN for greenhouse processes reported in the literature, it is not clear to which extent the reported models carry an inherent variability that could influence the results. An exception is given in the work by Ehret et al. (2008) and Ehret et al. (2011), who carefully report the meta-parameter tuning of neural networks models used to estimate plant-related parameters in greenhouses. However, it has been found in the literature that the comparison between ANN-models is commonly done on the basis of a single model run.

An example of such a model comparison is done by Boaventura-Cunha (2003). The authors compared a physical, an autoregressive and an ANN-based models by means of $rmse$ (root mean square error), but no explanation is given about the tuning of metaparameters. A similar situation occurs with the climate models reported by other authors (He and Ma, 2010; Fitz-Rodríguez et al., 2012; Trejo-Perea et al., 2009). In other cases, the comparison between models is made in a time basis i.e. one simulation against another from the same model (Trejo-Perea et al., 2009), although that is recommended against from the statistical point of view (Cumming et al., 2007).

The individual training runs in the present study were sensitive to the initial conditions of the weights, a fact mentioned by several authors (Ehret et al., 2011, 2008; Bishop, 1995). Indeed, for the climate prediction problem shown in Figure 15 on page 66, the traininig error achieved by 15 independent training runs spreaded along a range that made it difficult to differentiate between the models tested at all. While the search for an optimal neural architecture was not the aim of this study, this point was especially taken care of during the model design phase.

The choice of a suitable neural network architecture for a particular problem is an active area of research (Anders and Korn, 1999; Curry and Morgan, 2006; Saxén and Pettersson, 2006; Goodfellow et al., 2016). This choice includes the number and size of hidden layers in feedforward networks. In this work, the number of hidden layers was set to 1, and for this hidden layer, 3 number of nodes were tested (Section 3.3.1 on page 56). This approach was not intended as an exhaustive search for the best number of nodes. Authors like Ferreira et al. (2012) and Masters (1993) give a number of reasons why such an exhaustive search may not

be practical. Among these reasons is the above mentioned *noisy fitness evaluation problem*, which is "the fact that neural networks produce different results due to different initialization conditions even when everything else is kept fixed. This is why a single run is actually not enough to evaluate a topology" (Stathakis, 2009). Therefore, in this study, 15 different weight intialisations (originated by 15 random seeds) were tested for each model. As a comparison, Ehret et al. (2008), as well as Ehret et al. (2011) used 30 random seed initialisations in the design phase.

The 3 climate prediction models in this research (CM_{col}, CM_{ref} and CM_{both}) performed better with the same number of input and hidden nodes (ANN_2, see Figure 14b on page 58). This agrees with the results reported by Salazar et al. (2010b), who developed a temperature prediction model using neural networks, and found the best topology to have 11 input and 9 hidden nodes. A similar result is reported by Trejo-Perea et al. (2009), who found a combination of 4 input and 3 hidden nodes to yield the best results for estimating the energy consumption of a greenhouse. In both cases, the number of hidden nodes was similar to (while slightly lower than) the number of inputs used.

The same approach used for the selection of the hidden layer size was used to determine the number of previous steps to use as inputs in the climate prediction models. The fact that the error was bigger whenever 3 previous information steps were included suggests that the time constants governing the changes in temperature and relative humidity in the two research greenhouses lied between 5 and 10 minutes (previous steps 2 and 3, see Figure 16 on page 68). Pasgianos et al. (2003) also report 5 minutes as time constants for the greenhouse environment. The minute-scale of the time constants of the greenhouse air and construction materials was already reported by Bot (1989a).

In an autorregressive model of air temperature, Uchida-Frausto et al. (2003) tested inputs with a delay up to 20 minutes (4 time steps of 5 minutes each), and reported no improvement of the model with information (external weather) older than 15 minutes. Indeed, in most cases, data older than 10 minutes were not considered in their models. Also working with autorregressive models, López-Cruz et al. (2007) found a delay of 2 steps in the temperature (1 time step being 5 minutes) to be the best parameter. This correspond with the results found with the neural networks in the present research. However, they calculated an in-

dependent delay for each input signal, and used only one step for each of the four inputs tested (external weather conditions), although delayed 3, 2 or 1 steps. While the time spans found in both experiments are similar to those found in this research, the two models by the mentioned authors had the advantage of treating each variable independently, which might be a cause for the errors in the neural network models shown in Chapter 4: The same time steps were used for all the input signals.

The time scale used for the simulation plays a prominent role and should be investigated more carefully to allow for model-aided control. The literature on greenhouse climate modelling reports time constants for air temperature control loops which range from 30 minutes (Schmidt, 1996) up to 3 hours (Bot, 1989a). Depending on the mixture regime (convective or turbulent), it can take from 3 minutes (open greenhouse with strong winds) up to 30 minutes (closed greenhouse) for the air to homogenise into a uniform mixture (Roy et al., 2002; Fatnassi et al., 2002). The time needed for the air to homogenise inside the greenhouse, together with the high speed with which the ventilationpanes can fully open and close (faster than a time step) can play a role in the simulation errors found due to rapid changes in the actuators (Section 4.1.2 on page 71). Another related factor is the different diffusion velocity of sensible and latent heat fluxes (Schmidt et al., 2008): While the relative humidity can change very quickly when the ventilation opens, the temperature maintains the same trend for a longer time.

The inherent variability of the neural network models also influenced the input selection process. In particular, the models trained with a minimum and maximum set of inputs (labelled respectively $M_{(min)}$ and $M_{(max)}$, see Section 3.3.2) did not yield overall different results. This fact can be seen in Figures 15 and 16 on page 66 and on page 68, where the error bars suggest that that the climate predictions improved only marginally when more inputs were included (in some cases they even worsened). Since the error bars in the interaction plots represent the 95% confidence interval of the mean error for each model tested, they can be used to estimate the group differences, using the rules given by Cumming et al. (2007). According to these authors, the overlapping error bars suggest that there is no significant difference between groups, which was confirmed by an analysis of variance. The decision to select the simplest model candidate if their behaviour followed a common rule

in model selection, and a general scientific guideline, broadly known as "Occam's razor" (Anderson, 2008; Pearl, 2009).

While the overlapping error bars in Figures 15 and 16 hint at a similar performance of the models $M_{(min)}$ and $M_{(max)}$, the role of the remaining individual inputs was examined using Figure 17. It was found that the inclusion of more inputs into the model version $M_{(min)}$ and their exclusion from the model version $M_{(max)}$ were inconsistent between themselves and between models. The conclusion drawn was that they did not improve the behaviour of the model $M_{(min)}$ substantially. The selected inputs for climate prediction were limited to those in the simplest model version: air temperature, relative humidity, ventilation opening, thermal screen closure and 4 solar calculations regarding the time of day and year. This excluded several signals from the subsequent tests: the heating and cooling systems, as well as the leaf area index and the weather conditions outside the greenhouse.

Other authors report ANN-based models of grenhouse climate which rely on the external climatic conditions and the state of the actuators. For example, Fitz-Rodríguez et al. (2012) built a recurrent neural network to predict temperature and relative humidity using the weather conditions outside as well as the ventilation opening and fogging intensity inside the greenhouse, reaching a prediction error rate of $0.2°C$ for temperature and 2.3% for relative humidity, and regarding a prediction horizont of 10 seconds. Unfortunately, the authors do not report the number of iterations (previous steps) used. A key difference between the model reported by Fitz-Rodríguez et al. (2012) and the implementation shown in this study is the use of previous steps of variables not included in the output vector (hence the name: *NARX*, neural autoregressive models with exogenous inputs). In contrast, the neural networks in this study only had access to previous values of its own outputs. Fourati and Chtourou (2007) also used recurrent neural networks to predict the climate in a greenhouse, although their architecture (*Elman neural network*) does not feed back the outputs, but only the internal states. Therefore, this network could not be classified as an autoregressive model: The internal temperature and relative humidity were not present in the model inputs.

In both mentioned cases, the authors rely greatly on external weather and control inputs. This design follows the guidelines used for physical models of the greenhouse climate (Bakker et al., 1995). However, authors

like Seginer (1997) have pointed out the essential difference between the use of inputs in both approaches, as well as the importance of input reduction methods to ensure a good performance of the neural networks. The methods for input reduction they suggest (bottleneck neural networks, principal component analysis) underline the difference between inputs to the models and information content in terms of physical processes.

Indeed, authors developing so-called *blackbox* or *graybox* models (Oussar and Dreyfus, 2001) have reported that the external weather conditions are not always necessary to improve a greenhouse simulation model. For example, Salazar et al. (2010a), dealing with a model of greenhouse air enthalpy, compared groups of inputs including air measurements inside the greenhouse, weather conditions, greenhouse actuators and plant transpiration, and found that the external conditions (including the external air enthalpy) improved the model performance only marginally. Uchida-Frausto et al. (2003) also discussed the role of solar radiation in their autoregressive model:

> *"The influence of solar radiation can be seen to be relatively less important with respect to inside air temperature changes."*

They hint at the time constants as a potential cause:

> *"... solar radiation needs first to be absorbed by the plants, the soil and the construction, which subsequently exchange part of the heat gained in that way with the air through convection. This mechanism obviously damps the effect of short-term variations of the solar radiation level on the inside air temperature."*

When different mathematical models are available as candidates for a problem, not only the fit to the test data, but also the model complexity should play a role in model selection. This is called the *principle of parsimony* and is related to under- and overfitting models, as mentioned by Anderson (2008), who also mentions:

> *"As biologists, we think certain variables and structure must be in a 'good model' often without recognition that putting in too many variables and too much structure introduces large uncertainties ..."*.

The same author recommends using a measure of *information* to pick a model, particularly, the *Akaike Information Criterion* (AIC) or the *Bayesian Information Criterion* (BIC). This is the approach followed by López-Cruz et al. (2007) to select an autoregressive model. A comprehensive introduction to this Information-Theoretic approach can be found in Johnson and Omland (2004).

Unfortunately, the calculation of the AIC is not a straightforward task when dealing with ANN, due to the difficulty in the estimation of the (effective) number of parameters in the model (Murata et al., 1994; Anders and Korn, 1999). Murata et al. (1994) modified the AIC to a *Network Information Criterion* (NIC) which could be used to compare neural networks that can be considered submodels[1] of each other:

> *"The proposed NIC criterion measures the relative merits of two models which have the same structure but different number of parameters. In other words, when applied to neural networks, the criterion determines whether or not more neurons should be added to a network"*.

A careful network design should avoid the so-called *overfitting* of the model, that is, automatically incorporating noise as were it a structural part of the modelled process (Anderson, 2008). Overfitting is caused by an excesive number of parameters (a too complex model for the task at hand) (Curry and Morgan, 2006; Anderson, 2008). In the case of neural networks, this means too many hidden neurons (Murata et al., 1994; Masters, 1993). This implies that "growing errors in a validation dataset (during training) should be seen as an indication to reduce the network's complexity" (Anders and Korn, 1999), as the model would be overparameterized. For the three neural networks developed in this study, the learning curves in Figure 18 on page 70 show no signs of such a problem.

The concepts of *overfitting* and *underfitting* are closely related with the generalization error, also called test error (Goodfellow et al., 2016). In the end, we are interested in a model's capability to deal with data not seen before. ; Two of the models developed in this work were built

1 A submodel is generated through the elimination of a set of relational functions from a complete model (Pearl, 2009). For neural networks, this means the deletion of a number of connections between neurons, i.e. replacing their weights by a null constant value (Murata et al., 1994).

with data from a single greenhouse, the third one mixed data from both. The two single-greenhouse models opened the opportunity to make a test with data from each other. This can be thought of as an extended test dataset: The models were run on data from a different greenhouse (although very similar in many respects, see Section 3.1.1 on page 35).

It was expected that the models performed better with the test dataset taken from the same greenhouse than the training datset, however this was not always the case. There were noteworthy exceptions (Section 4.1.2 on page 71), where the test data was better modelled by either the *crossed* model (the model trained with the other greenhouse's data), or by the *combined* one. Still, no general pattern could be found: In some cases the combined model performed better than the corresponding ones (Figure 21 on page 74), while in others the opposite occurred (Figure 22 on page 74).

However, an overall error evaluation can only present a single facet of the models performance, hence the importance of the model runs presented in Section 4.1.2 on page 75. The selected test runs made it possible to identify that the models reacted differently, mainly with regard to the operation of the actuators. While the model CM_{col} tended to overreact more often to changes in the ventilation openings, the other two models did it more with changes in the thermal screen.

Nevertheless, the above mentioned results suggest that the use of data from different greenhouses for training and testing neural models could help to improve their performance. To our knowledge, this is the first work to test a neural greenhouse model with data from another facility, or to mix data from two greenhouses to build a single model. Future work should encourage the consideration of different data sources to the extent possible, even though this requires a lot of fine-tuning of data and model parameters, as was discussed regarding the model response to the actuators.

The errors due to the rapid operation of the actuators might be the biggest shortcoming of the climate models developed in this study, since an accurate relationship between actuators and controlled variables is a prerequisite for using the models in several control systems, for instance, those based in predicted scenarios to trigger control strategies on the run.

In this context, the work of authors like Linker et al. (2000), Fitz-Rodríguez et al. (2012) and Fourati (2014) will prove very valuable. The recurrent networks used by Fitz-Rodríguez et al. (2012) and Fourati (2014) could cope with the time issues presumably behind the inaccuracies described before. The modular approach used by both authors also allows to train independent neural networks for the greenhouse climate simulation and the controller itself. This architecture (as well as the nomenclature used by Fourati (2014): *direct, inverse neural model*) reminds of the models used in several works in the field of robotics, like those by Ziemke et al. (2005).

5.2 SIMULATION OF PHYTOMETRIC SIGNALS

The results from the model design listed at the beginning of Section 4.2.1 on page 80, when considered together, picture a clear pattern: The simulation of plant processes is more complicated than the simulation of the climate (even though the climate models in this study had a temporal dimension). Opossed to CM, all models in PM performed better with more complicated networks (more hidden nodes), reached larger MSE values, and benefited from the inclusion of additional inputs.

For leaf temperature and photosynthesis rate, the input selection process showed similarities between models trained from each greenhouse (and both of them). In the case of transpiration, the model trained with data from the collector greenhouse improved with every signal added as input, which led to the decision of including them all, with exception of the wind velocity (Table 12 on page 88).

The differences in the transpiration model trained in the collector greenhouse were not limited to its design: The simulation results were worse than those of the other two models. This model strongly underestimated the transpiration in both greenhouses (Figure 40 on page 94). The fact that the transpiration in the collector greenhouse was overall lower (due to the semi-closed operation resulting in high relative humidity) could account for the errors when the model was used to simulate the reference greenhouse (because the test data was partly outside the range of the training data), but not for the errors in the collector greenhouse itself.

A possible explanation could be given by the research conducted by Dannehl et al. (2014a,b). The authors identified significant differences in transpiration (among other plant indicators) between the collector and reference greenhouses during one of the cultivation years used in this research (2011), and related the differences with the control strategies used. But more interestingly, they reported differences in the content of secondary plant compounds, which points at physiological changes in the plants as a response to the climate control strategy. It could be possible to think of a physiological adaptation to the modified microclimate in the collector greenhouse, which would make it difficult to mimic the transpiration using only the information available in the inputs tested for the model PMb_{col}.

In the other two transpiration models (PMb_{ref} and PMb_{both}), the most noticeable input effects came from the leaf area index and the leaf temperature. The CO_2 concentration and the solar radiation showed an effect too, while not as marked as the former signals.

Regarding the selected inputs, it can be assumed that both the air and leaf temperatures already provide information about the solar radiation (as was discussed in Section 5.1). The influence of the leaf area index on transpiration is not as easy to explain, since the output variable is the transpiration rate per unit leaf area, and not the canopy transpiration. A possible explanation could be the correspondence between high values of LAI and high values of temperature: the plants are bigger in the summer, when the temperatures are higher than in the time near the transplantation date. The transpiration is also bigger in summer, due to the higher temperatures.

The plant transpiration strongly depends on the air temperature, and is regulated by the opening and closure of the stomata. Thornley and France report a strong air and canopy temperature dependency of at least seven parameters needed for the calculation of transpiration (using the Penman-Monteith equation), including the latent heat of vaporisation of water, the density of saturated water vapour and the CO_2 diffusion conductance (Thornley and France, 2007, p. 366, p 377). Moreover, Takakura et al. (1975) and Takakura and Fang (2002) define in their seminal work both the plant respiration and internal resistance to gas change as direct functions of the leaf temperature. The control of tran-

spiration developed by Hashimoto (1980) also relied on air humidity and leaf temperature, since it directly influenced the water stem flow.

In a series of more recent experiments, Schmidt (2002) reevaluated the resistance-based model of the leaves using a microscope to visually identify the stomata reactions to the climate, restating the strong dependency of the stomata conductance on leaf temperature and vapour pressure difference. Other interesting results reported by the author concern the time needed for the stomata to close, which had a minimum of 15 minutes; and the use of a 20 minute moving average filter, which revealed a strong correlation with the net photosynthesis of greenhouse tomatoes. These time constants might have had an effect on the simulation results by the photosynthesis models in this research, as will be discussed later in this section.

Using a multiple linear regression analysis, Jolliet and Bailey (1992) found the global solar radiation, the vapour pressure deficit and the air speed (all measured inside the greenhouse) to be the climatic factors with the strongest influence on transpiration. The authors point at a strong correlation between solar radiation, vapour pressure difference and air temperature, especially during the periods without greenhouse climate control (the experiments were conducted under a factorial design on climate set points). They found the Penman evapotranspiration model to have low accuracy under greenhouse covering, and linked this result to an overestimation of the solar radiation influence. They mention the constant stomatal conductance as a drawback of some of the tested models, and advise to use a dynamic model for it. Yet, they found it unnecessary to include the CO_2 concentration and air temperature in its calculation, as "these parameters do not affect directly the stomatal conductance or the transpiration of tomatoes". The latter statemen seems to collide with the previously mentioned results, and could be related to the time scale: The authors used transpiration measurements averaged over periods between 0.5 and 2 h. However, since the method for input evaluation used by the authors was multiple linear regression, it is more likely that the discrepancy was associated with the correlations in the model inputs reported in the same paper.

Unlike transpiration, the 3 neural networks developed as photosynthesis models (PMc_{col}, PMc_{ref} and PMc_{both}) used the same input variables: CO_2 concentration, leaf temperature and transpiration rate (in

addition to air temperature, relative humidity and the solar calculations in the minimum set). There are two ways of looking at the leaf temperature as an input to the photosynthesis model. Firstly, in a purely data-oriented aspect, the leaf temperature (and air temperature) follows a daily cycle, and thus correlate with the solar radiation. Secondly, from a physiological point of view, there is evidence of a temperature dependency of the photosynthesis. In particular, the enzimatic activity of $RuBisCO$ depends on the leaf temperature and directly affects CO_2 assimilation (Farquhar et al., 2001; Von Caemmerer and Farquhar, 1981). The fact that the photosynthesis simulation profited from the inclusion of the transpiration as an input can be explained by the importance of the water content in the leaves (Haefner, 1996; Farquhar et al., 2001), but also by the strong dependence of photosynthesis upon the stomatal conductance (Schmidt, 2002).

Katsoulas et al. (2001) showed the variation of the stomata conductance as a function of vapour pressure deficit by means of fogging on a rose culture. While the authors did not find big differences in the transpiration rate, the low VPD achieved during fogging favoured an increased stomatal conductance, even at noon. This points at water stress avoidance and could help improve the photosynthesis during the periods of high solar radiation.

The fact that the stomatal conductance is a key factor to the photosynthesis has been widely studied by Wong et al. (1979) and Farquhar and Sharkey (1982). Also, plants growing under altered VPD-levels have been found to modify their stomatal conductance lastingly as an acclimation response, and the thus increased stomata sensitivity to CO_2 has been linked to increased net assimilation rates (Kawamitsu et al., 1993; Talbott et al., 2003).

In this respect, the differences in relative humidity (and thus vapour pressure deficit) between the two experimental greenhouses used in this research could account for the differences in photosynthesis reported by Dannehl et al. (2014b), due to the semi-closed operation of the collector greenhouse. As was described in Section 4.2.2, and particularly on Figure 42 on page 96, the photosynthesis rate was underestimated by all three models. This simulation error, more marked at higher rates of photosynthesis, might be a consequence of the models being unable to properly model the collector greenhouse (where the highest photosynthesis

rates were recorded). However, considering the fact that all three models underestimated high photosynthesis values and overestimated low ones, it is likely that the error was also influenced by the sigmoid activation function used in the neural networks (Section 2.2.1 on page 25).

In the big picture, the single over- and underestimation errors added up to a filter-like effect of the whole simulated signals, more marked in the photosynthesis, but also present in the transpiration. Many fluctuations observed in the measured signals were automatically smoothed out by the neural networks. Yet, this cannot be traced back to the use of a low-pass filter to preprocess the data (Section 3.2.1 on page 48), since the same procedure was applied to the train, validation and test datasets.

Aside from over- and underestimations, one factor that contributed to the apparent filtering of the simulations of transpiration and photosynthesis was the use of theoretical solar inputs, which did not take quick changes (like the sudden presence of clouds) into account.

Model inputs that change quickly can lead to simulation errors, if a single data point is considered at a given time. As an example, consider the effect of the wind velocity as an input signal, which improved the model performance in some runs while worsening it in others, thus pointing at a random effect more than a real, clear improvement. In these cases, a single 5-minutes time step does not reflect the whole dynamics of the plant processes being simulated.

It is worth to remember that PM did not incorporate any chronological structure. Their temporal static nature could speak against a smoothing of the network's output, since every data point is taken as a training or simulation unit, with no reference to the signal values before and after it, neither during training nor in the simulations.

Sánchez et al. (2012) reported a similar result when dealing with autoregressive models of the transpiration of a tomato crop. Although they report this result as a *delay*, the simulation runs by the authors resemble those in this work (Figures 44 and 45). Rather than a change in phase, it appears to be damped. The authors refer to the work of Gázquez et al. (2008), who identified the phenomenon as *hystheresis* between transpiration and vapour pressure difference. They point at an imbalance between the transpiration rate and the root re-absorption (since the VPD was strongly influenced by a fog system).

Indeed, a possible explanation for the smooth outputs from PM (both in transpiration and photosynthesis) is given by the time period needed by the plants to react to a changing environment. Different authors had reported that the stomatal conductance and net CO_2 assimilation involve reactions to phenomena occurring approximately 15 to 20 minutes before (Pearcy et al., 1985; Jones, 1998; Schmidt, 2002). The relationship between transpiration and stomata conductance is also highly dynamic, with the conductance being controlled by the transpiration rate itself, rather than the humidity content of the air (Monteith, 1995). Conversely, this control mechanism is reported to gain importance as the air movement breaks down the resistance of the boundary layer next to the leaves (Jones, 1998), a situation less common under greenhouse conditions (and even more in semi-closed greenhouses) than in open-field agriculture.

One side effect of the aforementioned smoothing effect was that the models partly compensated the presence of errors in the networks' inputs, thus making the simulations more robust. This effect was more noticeable when the transpiration and photosynthesis were simulated with already simulated phytomonitoring signals. In these cascade simulations (Figure 46 on page 103), some inaccuracies introduced by the first and second phytometric signals models (leaf temperature and transpiration rate) were bypassed by the second and third ones (transpiration and photosynthesis rates).

The possibility to connect in cascade the three submodels in PM made it possible to use them in two ways:

- Taking advantage of measured values of leaf temperature and transpiration rate whenever they were availabe, in order to improve the simulations of transpiration and photosynthesis.
- Using simulated values of leaf temperature and transpiration rate when no measurements were present, in order to provide continuous estimations of transpiration and photosynthesis.

The main benefit of the second configuration is that it is able to give estimations even if the complete plant measuring system fails, which is an important safety feature towards plant-oriented control stategies. This simulation scheme was demonstrated by means of an example featuring a complete outage of the sensor system in section 4.2.2 on

page 102, having in mind that a if a greenhouse control system is expected to run continuously, the reliability of the sensors is crucial. It has been pointed out by several authors (e.g. Jones (2004); Katsoulas et al. (2001); Villarreal-Guerrero et al. (2012b)) that the availability of on-line measurements is a prerequisite for the development of irrigation control strategies based on plant transpiration. This assumes the presence (and functionality) of a minimum set of sensors. Concretely, the models shown here (PMa, PMb, PMc) needed at least measurements of air temperature, relative humidity and CO_2 concentration inside the greenhouse, as well as temperature outside it.

Models, which are developed to estimate the output of a sensor or measurement system are also sometimes called *soft-sensor* or *virtual sensor*. Sánchez et al. (2012) developed a virtual sensor as a low-cost, low-maintenance alternative to a microlysimeter continuously weighing a small row of 6 tomato plants. Their virtual sensor, like the models in this study, depends on the availability of only a limited number of sensors (air temperature, relative humidity and solar radiation) in order to calculate the evapotranspiration of a tomato crop.

Sánchez et al. (2012) also mention two important uses of virtual sensors: The direct replacement of the physical sensor (thus reducing hardware costs), and the simultaneous comparison of simulated and measured signals to ensure the vailidity of the measurements. In the latter case, i.e. when a measurement and a model are expected to run in parallel, it is necessary to have a means to tell valid from invalid measurements. Therefore, further work along the line of model-aided greenhouse control should include the development of online failure detection systems to pick (or calculate) the best possible value of the phytometric signals to be used by the controller. Fortunately, the time response of the plants to the climate should facilitate the detection of such errors in time windows under 5 minutes.

Whether measured with a physical (Schmidt and Exarchou, 2000) or virtual sensor (Sánchez et al., 2012), or calculated via an energy balance (Harmanto et al., 2005), the most widespread use of the crop evapotranspiration has been to define an irrigation schedule. The determination of the moment and amount of water to be applied is a topic long studied in open-field agriculture (Allen et al., 1998). However, difficulties to calculate the reference evapotranspiration (ET_0) under greenhouse con-

ditions (Villarreal-Guerrero et al., 2012b; Fazlil-Ilahil, 2009) have led to new approaches to irrigation control.

The rationale behind the above mentioned irrigation methods is to guarantee that the water potential in the soil does not fall under a minimum value, in order to avoid stress and thus preventing stomata closure. An extreme case of this approach is given by hydroponic systems, where the problem of water availability in the soil is replaced by other concerns (nutrient availability, toxicity, oxygenation, spread of pathogens).

One example of a different method is given by Steppe et al. (2008), who developed an irrigation control based on online measurements of the water potential in the stem of apple trees, together with sap flow. Their work uses the water potential to determine the irrigation moment, defining the amount of water to apply by the integral sum of the sap flow.

Jones (2004) has criticized plant-based techniques that rely on the water potential in the plant by itself:

> it is apparent that [...] the favoured way to use plant water status is actually as an indicator of soil water status; this negates many of the advantages of selecting a plant-based measure!

Nonetheless, one core point of the controller presented by Steppe et al. (2008) is the link between irrigation and photosynthesis, by means of a trigger value for watering the trees based on the maximum photosynthesis rate. While the authors report this link as a single trigger value (which had to be adjusted anyway), the logic behind is intrinsically different from the mentioned soil-water approach: The irrigation should be controlled to ensure a proper photosynthesis (and not transpiration itself).

Indeed, an interesting line of research regarding model-aided control of greenhouses could shift the focus from individual controllers to an integrated system linking the actual important control variable: The production of edible parts. Maclean et al. (2012) point out that the plant models needed for control purposes need to be simple and focus on production to allow for plant-based control, instead of control of environmental variables. In this point, they refer to what Marcelis et al. (1998)

define as *descriptive models*: Models with few states and low computation time. The latter author also stresses the importance of harvestable organs (or dry matter production) in the simulation models.

In this respect, techniques like ANN can prove valuable because of their capability to integrate information from different sources and to run quickly once trained. One challenge is, however, to find appropriate ways of mixing short-term control actions with mid-term plant processes, i.e. to draw and execute control commands from projected photosynthesis and dry matter accumulation. In their survey, López-Cruz et al. (2014) found only few neural models that take aim at control, being the most of them predictive models of yield or quality in a weekly time scale. Such time scale allows to correct control strategies (e.g. day/time temperature set points), but fails to link the plant response directly to the actuators in the sense of the now classical *speaking plant approach* (Udink ten Cate et al., 1978).

5.3 LONG-TERM PREDICTION

The experiments shown in Section 4.3 on page 106 lay out a connection scheme that enables to link the control actions of the greenhouse with the expected plant response. Additionally, the fact that CM was designed as a short-term prediction model allowed to chain predictions and thus create long-term predictions of the plant response itself, in intervals up to 30 minutes.

The interconnected system was intended as an explorative work towards a model of the greenhouse dynamics that was suitable for plant-oriented control strategies. A key element of the coupling was the long-term prediction of climate. The jump from the one-step to the long-term prediction of climate presented itself as a natural research question, dealing with only a single model in a recursive manner.

The recursive prediction of greenhouse climate has often taken the form of autoregressive models. However, the research results found in the literature most commonly consider the recursion as a means to achieve a one-step prediction, i.e. the autoregressive nature of the models lies in the past: Previous measurements are used to estimate a single point in the future. Therefore, it is difficult to make a one-to-one comparison

of the models in the present study with those in the literature. For example, Uchida-Frausto et al. (2003) used up to 4 previous steps (5 minutes each) of information to build ARX and ARMAX models that predicted the inside temperature 1 step in the future. Their seasonal models achieved reported accuracies as high as $r^2 = 0.988$ (ARMAX) and $r^2 = 0.989$ (ARX) (both cases in springtime). These values could be located between the second and third simulated steps in Figure 53 on page 116. Two main differences between the models were that those in the present work were already part of an LTP, and that they refer to year-round cultivation periods (instead of seasonal models). The same authors (Uchida-Frausto and Pieters, 2004) incorporated a neural network after the ARX model, and achieved a year-round mean absolute error of $1^{\circ}C$, which represents a bigger error than that of the models in this study, even after 6 prediction steps. Interestingly, the authors report that the neural network helped to capture the non-linear effects of the ventilation, a problem that could not be solved in this work (as will be further discussed in this section). Similar results were reported in a tropical climate by Patil et al. (2008): For a seasonal (mid-summer) model, their model reached values of $r^2 = 0.973$, $r^2 = 0.976$ and $r^2 = 0.968$ for an ARX, ARMAX and NNARX architectures, respectively.

One difficulty to make long-term predictions with models incorporating exogenous inputs (for instance the weather conditions outside a greenhouse) is the lack of future values for those variables. It is unclear which methods were used to address this problem in some research regarding recursive simulations that make long-term prediction of greenhouse climate (Boaventura-Cunha, 2003; Dariouchy et al., 2009). The long-term predictions in this study were possible because the only measurements involved were those of the simulated variables, being all exogenous inputs suitable to be calculated as a sole function of the time (solar coordinates and leaf area index).

There was one exception to the previous statement. The only measurement involved in the long-term predictions was the CO_2 concentration used as input to PM. Since the simulations in Section 4.3 were run offline, it was possible to use the already-known measured values of this signal as model input. There are several alternatives to adapt this signal to an online control model: Firstly, Salazar-Moreno et al. (2011) has shown that it is possible to forecast the CO_2 concentration itself; secondly, being the CO_2 concentration a control variable, a set-point value

can be obtained from the automation system; and finally, for a limited number of prediction steps, it is possible to make the assumption that the concentration either remains constant or that it follows a foreseeable trend.

Apart from the autoregressive schemes mentioned, some research has focused on climate prediction models which do not rely on recursion, and therefore directly relate input variables in $t = 0$ with the predicted output. Among them is the work by He et al. (2007), who built a neural model to predict the temperature and relative humidity inside a greenhouse in a 30-minutes OSP as a function of the current external climate. For air temperature, the authors report an $rmse = 0.8°C$, a value similar to the fourth or fifth predicted steps in Figure 53, which correspond with a prediction of 20 to 25 minutes. For relative humidity, the authors report an $rmse = 1.1\%$, which could only be achieved in the OSP of the models in this work. One explanation for the high accuracy in the prediction of relative humidity can be given by the small dataset (300 records) used.

A last example from the literature helps to give context to the results in Figure 53: A dynamic, mechanistic model to estimate the climate inside a greenhouse (without making predictions in time) was presented by Guzmán-Cruz et al. (2009). The calibrated model simulated temperature and relative humidity and got correlation coefficients as big as 0.9187 ad 0.9332, respectively. The models in this study achieved similar values even after 6 simulation steps of temperature or 5 of relative humidity.

While the former examples show that the error measures describing the climate predictions in this work make them comparable with other models in the literature, it is clear that a single value cannot capture the particular behaviour and responses of the models. The overall statement made by a numerical measure of error can be useful to compare models created and tested under similar circumstances, and help pick and discriminate between them. However, a numerical measure alone is unable to tell details about *how* and *when* the models can be useful, nor can it hint at problems and improvement areas.

In Section 5.1 it was pointed out that a specific problem of the climate prediction model was related with quick changes in the greenhouse

actuators (ventilation opening and thermal screen closure) (see also Figure 27 on page 79). For the long-term predictions, this meant that a simulation train was prone to sharp deviations if it started with a big error in the one-step prediction. It was observed that the operation of the thermal screen affected the prediction of air temperature more than that of relative humidity. Normally, the opening of the thermal screen occurred in the early morning, and the closure in the late evening, times at which the relative humidity was very high, with values approaching 100% and little variations. The high and relatively stable values of humidity could therefore account for the smaller errors observed in its prediction.

Being operated mostly once a day, the data regarding changes in the thermal screen closure were rather scarce in comparison with the complete amount of data records available. In this context, Linker et al. (1998) hint at a possible cause for the poor model learning of such situations: Data sparsity. The authors advise to split the neural models to include only one simulation output (that is, separate neural networks for air temperature and relative humidity) and stress the dimensionality reduction by means of dropping input variables which contribute only marginally to the output at hand. In contrast, Fitz-Rodríguez et al. (2012) suggest to directly gather more data through physical tests. Another possibility to deal with sparse training data is given by data augmentation techniques, which allow to make up additional samples by slightly modifying the already available data (Goodfellow et al., 2016; Schmidhuber, 2015).

An unexpected result was found in the long-term prediction of photosynthesis, which was not particularly affected by the aforementioned errors. This observation was more remarkable in the light of the errors being present simultaneously in several model inputs: What started with a sudden change in the actuator inputs to CM propagated thereafter to the simulated leaf temperature and (to a lesser extent) to the transpiration rate, but was barely noticeable in the simulated photosynthesis. No clear explanation for this phenomenon could be found, and thus remains an open research question.

There are two sides to the long-term prediction of photosynthesis not reflecting the errors due to the operation of the actuators. On the one hand, the predictions gained in robustness, being less prone to error propagation and thus suitable for use even with noisy inputs. On the other hand, this robustness can also be seen as a lack of sensitivity to the

input signals. Depending on the intended application for the simulation model, one of these facets may prevail and turn into an advantage or a drawback.

Two main application areas of the coupled models can be identified: As a plant virtual sensor and as an element in greenhouse control. The use as a virtual sensor that either replaces or runs in parallel with physical measurement systems was already mentioned in the previous section. A recent application of such a model, to monitor the overall plant responses and implement and trigger warning messages to the grower was recently reported by Tsafaras and de Koning (2017). One example of virtual sensor application could be the use of a sophisticated hardware measuring system to gain data on-site, train a neural network and then replace the hardware by the trained model.

Conversely, the integration of phytometric signals in the greenhouse control loop demands the integration of several technical elements and can been addressed in the following ways:

- The expected plant response can be used to establish the set point of the control signals already present in the system. This process can be static (Steppe et al., 2008) or dynamic (Aaslyng et al., 2003). In both cases, the controller itself does not need to be modified as it only receives a modified set point. If the simulated plant response lays in the future, the system can perform predictive control.

- From the prediction of reasonable control scenarios, an optimum control action can be selected on the fly. One way to implement this approach is the so-called *action selection* (Goodfellow et al., 2016; Schillaci et al., 2012; Ziemke et al., 2005), where the operation of the actuators can be directly selected. Another one is shown by Aaslyng et al. (2003) too, in the form of a set-point selection based on the multi-objective optimization of photosynthesis and energy cost. In either case, the forecasted scenarios must be continuously evaluated by a parallel module, responsible for the decision taking.

- Neural control systems are able to directly incorporate phytometric signals as inputs, if they are included in the direct model of the

plant[2] In this approach, the modelled *plant* can be used to offline train the neural controller (Fourati, 2014; Nguyen and Widrow, 1990b).

Research works like those presented by Aaslyng et al. (2003), Salazar-Moreno et al. (2011) and Fourati (2014) set guidelines towards integrating control systems to be used in greenhouse production. They empahsize the whole control process, from the actuators until the plant response, and back to the control actions, as the latter represent the levers available to operation.

Another issue that needs to be tackled towards plant-oriented control is the model portability, i.e. the capacity to use a trained model in one greenhouse to model another one. When using ANN, Salazar-Moreno et al. (2011) and Linker and Seginer (2004) suggest the use of pre-training (or re-training), that is to say, to periodically train the networks after a sensible amount of new data is gathered on-site. Another possibility could be the use of other training algorithms (apart from backpropagation), which include the so-called *reinforcement learning* (Goodfellow et al., 2016; Schmidhuber, 2015).

⋆

Many of the subject matters discussed in this chapter (and particularly in the last part of it) are active research topics in engineering and industry, with a pushing development that follows the recent increase in data-driven applications in many different areas. These applications often involve the use of ANN of greater complexity than those shown in this work (keyword: *deep learning*). Therefore, the models developed and analysed in the present study can be taken as a starting point of research lines that will need the commitment of interdisciplinary workteams to allow the greenhouse horticulture to profit from the current progress in fields like computer science and machine learning.

2 The term *plant* has the meaning used in classical control theory and engineering. In the particular case at hand, it comprises the greenhouse climate, the technical equipment as well as the plants themselves.

6

CONCLUSIONS AND OUTLOOK

The aim of the present work was the design and development of neural networks for greenhouses. Particularly, ANN were to be used to implement models of the internal climate of a tomato greenhouse, as well as the plant's reactions to that climate. These facets of a greenhouse reflect the main objectives defined in Section 1.2 on page 2:

1. To implement predictive neural models of the greenhouse air temperature and relative humidity in a time scale of minutes.

2. To implement neural models of the following phytometric signals:

   - Leaf temperature
   - Transpiration rate
   - Photosynthesis rate

In pursuing these aims, it was expected to shed light on a series of transversal research questions that are important for the practical usage of neural models in automatic greenhouse control. Therefore, special emphasis was put on the model design process, as well as practical issues like the size of the modelling time steps, the modularity and connectivity of the models and the use of standard sensors and signals as model inputs.

This chapter gives the conclusions of this thesis and points at the insights gained about the research questions. It also includes research directions that can further this work as well as general recommendations for the field.

6.1 CONCLUSIONS

In addressing the first main objective, the air temperature and relative humidity were considered as model outputs. One key difference with other climate models reported in the literature is that this work assumed the presence and reliability of sensors of those both variables. By contrast, other authors have reported models that calculate the temperature and relative humidity as a function of the outside climate alone. While the latter is a very intuitive point of view, especially when dealing with mechanistic models, it oversees two facts: a) that, in the short term, these variables already contain information about their own change; and b) that the sensors of temperature and relative humidity are cheap and robust, and most (if not all) commercial greenhouses use them regularly. The first point, that these variables give information about their own trajectories, is considered by all forms of autorregressive models. The models in this thesis also profited from this technique, by including previous measurements as model inputs. It was tested how many past measurements to use, finding that 5 to 10 minutes of previous information (current, plus 2 previous steps) would help to improve the simulations. The second point, the widespread use of temperature and relative humidity sensors, allowed to test the convenience of the use of an external weather station, which are not always available near the greenhouses. It was shown that the external weather did not clearly contribute to improve the short-term climate predictions.

In the same line of thought that used the current climate measurements as model inputs, the theoretical coordinates of the sun and the current state of two greenhouse actuators (thermal screen and ventilation opening) were set to be model inputs. The assumption was that these values are easily available and cheap.

The selection of inputs for the climate prediction model considered the additional value given by each candidate signal. A common misconception when dealing with neural models is that they *rank the inputs*, or *automatically select the inputs*. This was shown not to be the case: the models trained with all available inputs had a similar (in some cases even worse) performance than the models with less inputs. Particular care was put into this input selection procedure, and it was shown that a great deal of variability can be traced back to the random initialisation of the

network's models. This fact is commonly overlooked (or not reported) in published research, difficulting model selection: it is not always clear if an apparent improvement of a model was due to a random effect.

However, the selection of a number of inputs to serve as the starting point for model design was itself an assumption of this research and should be explored in further detail.

The selected models were used to run simulations that showed their behaviour on the time scale, i.e. one prediction after the other in a run chart. This allowed to detect that the sudden change of the actuators (ventilation opening and thermal screen) would bring about simulation errors in temperature and relative humidity. These errors would remain concealed if the models were tested using numerical measures of error only.

The presence of simulation errors due to the operation of the actuators represented a major problem for the climate prediction models, since they were intended to be used for control of those very variables.

Apart from the mentioned issues, the models achieved a very good performance, which is particularly notorious for a model applied over long periods (three years). In fact, most published research deals with short periods of time, or with a different time resolution (for example, time step of one day or one week). This allows to conclude that neural models of climate can be used in practical applications if they are trained with sufficient data, care is taken on the input selection and the role of the actuators is accounted for.

The experiments in this thesis showed that the simulation results are very sensitive to the tuning of the network's metaparameters. One consequence is that a direct comparison of models in the literature is difficult whenever these metaparameters are not properly reported. It remains open to develop ways to address this issue in a scientific context. The use of software suits that obscure parts of the model development process also contributes to this difficulty.

The second aim of this thesis was the implementation of models for three phytometric signals: leaf temperature, transpiration rate and photosynthesis rate. Unlike the climate prediction model, the phytometric signals were simulated separately, that is, three independent models were developed and coupled together later on.

It was shown that the leaf temperature can be simulated with reasonable accuracy by using only climatic factors (and leaf area index) as model inputs. The leaf temperature could be subsequently used as input for the transpiration model. Finally, the photosynthesis model would profit from having leaf temperature and transpiration rate among its inputs. This cascading was shown to be a flexible and robust configuration by means of an example consisting on a sensor outage.

As with the climate prediction models, it was found that the solar radiation measurements taken outside the greenhouse did not greatly improve the simulation of the phytometric signals. It was argued that the information provided by the radiation would be implicitly present in inputs already. Neural networks do not cope with information redundancy or correlated inputs automatically.

In all experiments, three model versions were developed in parallel: one with data from each grenhouse and a third one using mixed data from both greenhouses. Although both facilities were very similar in construction, cultivation techniques and had the same tomato crop, these tests showed that a model from one greenhouse could be useful to model another one. The possibility to train a neural model using a greenhouse's data, and subsequently run it with other greenhouse's sensors could greatly help to use neural networks in commercial automation computers. While these results are not conclusive (because of the similarity between greenhouses), they encourage further research in a topic that is not yet addressed in published research, to our best knowledge.

Lastly, a connection scheme was proposed to link the prediction of the air conditions and the simulation of the phytometric signals. This connection scheme, presented in Section 4.3, allowed to chain climate predictions on the one hand, and to use these predictions to estimate the plant responses, on the other hand. The connection scheme allowed to trigger *long-term predictions* of the greenhouse climate several steps ahead, leading to the corresponding long-term prediction of the phytometric signals. The climate trajectories tended to drift strongly after 2 to 3 predicted steps (about 20 minutes in the future), an effect even more noticeable when the ventilation opening and the thermal screen were operated. The long-term prediction of the climate has therefore potential for improvement. Surprisingly, the photosynthesis signals deteriorated (drifted) the least, even after several prediction steps and despite the

fact that they were simulated with already simulated phytometric signals (Section 4.2.2). This was an unexpected result, and should be further investigated, as it could support the idea of a photosynthesis-based control system.

6.2 OPEN QUESTIONS AND FUTURE WORK

A number of questions emerged from this work, including the following:

- How to link, incorporate the effect of the motor commands (actuators) on the climate prediction? How to include the heating and cooling systems as well?
- Which time delays are involved in the causal relationships acting upon the greenhouse climate? Are they the same for all input variables?
- How to determine the minimum set of sensors needed to reliably predict the greenhouse climate?
- Which input signals can be dropped from a neural climate model due to redundant information? Which input signals can be dropped from a neural model of phytometric signals? How does this redundancy depend on the time scale?
- How does a neural model perform when tested in a different facility? How does it perform with another crop? Is it possible to port neural models at all?

In addition to the modelling questions, further work should contemplate climate control. A previous condition is, however, to improve the simulations accounting for the effect of the actuators. One direction would include the use of recurrent neural networks with independent time steps. This design direction should consider the use of deep neural networks, particularly the so-called long short-term memories (LSTM).

The long-term prediction of climate and phytometric signals could be the basis for an action selection algorithm based in control trajectories: if an specific control action is carried out at a certain point, the effect on the plants after a number of time steps could be estimated before it

occurs. The selection of a motor action based on simulations could then be implemented as a plant-based climate control technique. The rationale behind this is known as *receding horizon* in control theory.

The input selection method should start with a cross-correlation analysis and it would be interesting to look at the information content of each input signal, from the point of view of information theory. This line of research could also include developing causal models and bayesian networks.

The use of previous steps of data, used in this thesis to improve the climate predictions, could also be tested for the photosynthesis models directly, if predictions of this variable are aimed at. In any case, testing the number of previous steps independently for each input signals could improve the model performance. However, a trade off must be considered, as the combinations of parameters can quickly multiply and make the model design prohibitive.

A very important area for practical applications is the portability of the models. Research along this line should include data collection from as many greenhouses and crops as possible, and testing the models with different situations. It would be interesting to construct a data sharing project, or the integration of a big joint dataset. Another project would be the development of adaptation techniques that allow a model to be installed in a greenhouse and adjust itself to it over time. One starting point would be re-train the models, for example once a month. A more sophisticated idea would be the use of adaptive training algorithms that could learn online (as opposed to the backpropagation algorithm). Lastly, pre-trained networks have become popular in other application areas (mainly image recognition) and pre-training (partial training) could also be adapted to greenhouse applications.

Lasty, ANN could support the use of models based on expensive and sophisticated plant measuring devices. As an example, it would be possible to develop custom-made models after a number of measurements were taken on production greenhouses. Such an approach could help to provide the growers with a feedback from the plants, without the need for a costly investment in hardware.

6.3 THESIS CLAIMS

Directly related to this research

1. Artificial neural networks can cope with the complexity of biological systems, and can thus be used to build models if enough measurements are available.
2. Neural networks are very sensitive to the calibration of metaparameters. Neural models can hardly be compared without proper information on metaparameter tuning.
3. The prediction of the greenhouse climate can greatly profit from the use of past measurements.
4. The leaf temperature can be used to improve the simulation of transpiration rate. Similarly, leaf temperature and transpiration rate can improve the simulation of photosynthesis rate.
5. Reliable long-term predictions of greenhouse climate would allow for predictive plant-based control strategies.
6. Neural networks can profit from non-causal relationships between signals and can thus help to save costs due to measuring devices.
7. Within boundaries, neural models can be trained in one greenhouse and used to model the climate and plant processes in another one.

Additional claims

8. Neural models will not become widely accepted until a critical number of growers and research stations share their data to develop joint models.
9. The ubiquity of internet services can help to integrate data from different greenhouse facilities.
10. Recent developments in neural networks (architectures, algorithms, activation functions) will greatly impact in their usage in agriculture and horticulture.

11. Interdisciplinary teams of plant and data scientists are needed to accomplish the transfer of neural technologies to the horticultural sector.

⋆ ⋆ ⋆

BIBLIOGRAPHY

Aaslyng, J. M., Lund, J. B., Ehler, N. and Rosenqvist, E. (2003). Intelligrow: a greenhouse component-based climate control system. Environmental Modelling & Software *18*, 657–666.

Akutsu, M., Sunagawa, H., Usui, T., Tamaki, M., Taniai, N., Hirata, M., Kaiho, A. and Takakura, T. (2015). Non-destructive, real-time, and automatic measurement of transpiration from a plant canopy stand. Journal of Advances in Agriculture *5*, 677–683.

Allen, R. G., Pereira, L. S., Raes, D., Smith, M. et al. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56, vol. 300,. FAO, Rome.

Anders, U. and Korn, O. (1999). Model selection in neural networks. Neural Networks *12*, 309 – 323.

Anderson, D. R. (2008). Model Based Inference in the Life Sciences: A Primer on Evidence. Springer Nature.

Avissar, R., Avissar, P., Mahrer, Y. and Bravdo, B. A. (1985). A model to simulate response of plant stomata to environmental conditions. Agricultural and Forest Meteorology *34*, 21 – 29.

Baille, M., Baille, A. and Laury, J. C. (1994). A simplified model for predicting evapotranspiration rate of nine ornamental species vs. climate factors and leaf area. Scientia Horticulturae *59*, 217 – 232.

Bakker, J. C., Bot, G. P. A., Challa, H. and Van de Braak, N. J., eds (1995). Greenhouse Climate Control - an integrated approach. Wageningen Press.

Bender, E. A. (2012). An introduction to mathematical modeling. Courier Corporation.

Bennis, N., Duplaix, J., Enéa, G., Haloua, M. and Youlal, H. (2008). Greenhouse climate modelling and robust control. Computers and Electronics in Agriculture *61*, 96 – 107.

Berg, E. and Kuhlmann, F. (1993). Systemanalyse und Simulation für Agrarwissenschaftler und Biologen. Ulmer.

Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.

Blasco, X., Martínez, M., Herrero, J. M., Ramos, C. and Sanchis, J. (2007). Model-based predictive control of greenhouse climate for reducing energy and water consumption. Computers and Electronics in Agriculture *55*, 49–70.

Bly, B. M. and Rumelhart, D. E. (1999). Cognitive science. Academic Press.

Boaventura-Cunha, J. (2003). Greenhouse climate models: An overview. In Proceedings of the EFITA Conference pp. 823–829, EFITA.

Bot, G. and van Dixhoorn, J. (1978). Dynamic modelling of greenhouse climate using a minicomputer. Acta Horticulturae *76*, 113–120.

Bot, G. P. A. (1989a). Greenhouse simulation models. Acta Horticulturae *245*, 315–325.

Bot, G. P. A. (1989b). A validated physical model of greenhouse climate. Acta Horticulturae *245*, 389–396.

Bot, G. P. A. (1993). Physical modelling of greenhouse climate. In The Computerized Greenhouse, (Hashimoto, Y., Bot, G. P., Day, W., Tantau, H.-J. and Nonami, H., eds), chapter 2, pp. 51–73. Academic Press Boston.

Boulard, T. (2012). Recent trends in protected cultivations-microclimate studies: a review. Acta Horticulturae *957*, 15–28.

Boulard, T., Feuilloley, P. and Kittas, C. (1997). Natural ventilation performance of six greenhouse and tunnel types. Journal of Agricultural Engineering Research *67*, 249 – 266.

Boulard, T., Kittas, C., Roy, J. C. and Wang, S. (2002). Convective and ventilation transfers in greenhouses, part 2: Determination of the distributed greenhouse climate. Biosystems Engineering *83*, 129–147.

Bronchart, F., De Paepe, M., Dewulf, J., Schrevens, E. and Demeyer, P. (2013). Thermodynamics of greenhouse systems for the northern latitudes: Analysis, evaluation and prospects for primary energy saving. Journal of Environmental Management *119*, 121–133.

Challa, H. and van Straten, G. (1993). Optimal diurnal climate control in greenhouses as related to greenhouse management and crop requirements. In The Computerized Greenhouse, (Hashimoto, Y., Bot, G. P., Day, W., Tantau, H.-J. and Nonami, H., eds), pp. 119 – 137. Academic Press Boston.

Cheng, B. and Titterington, D. M. (1994). Neural networks: A review from a statistical perspective. Statistical science *9*, 2–30.

Cumming, G., Fidler, F. and Vaux, D. L. (2007). Error bars in experimental biology. The Journal of Cell Biology *177*, 7–11.

Curry, B. and Morgan, P. (2006). Model selection in neural networks: Some difficulties. European Journal of Operational Research *170*, 567 – 577.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems *2*, 303–314.

Dannehl, D., Josuttis, M., Huyskens-Keil, S., Ulrichs, C. and Schmidt, U. (2014a). Comparison of different greenhouse systems and their impacts on plant responses of tomatoes. Gesunde Pflanzen *66*, 111–119.

Dannehl, D., Josuttis, M., Ulrichs, C. and Schmidt, U. (2014b). The potential of a confined closed greenhouse in terms of sustainable production, crop growth, yield and valuable plant compounds of tomatoes. Journal of Applied Botany and Food Quality *87*, 210–219.

Dariouchy, A., Aassif, E., Lekouch, K., Bouirden, L. and Maze, G. (2009). Prediction of the intern parameters tomato greenhouse in a semi-arid area using a time-series model of artificial neural networks. Measurement *42*, 456–463.

Dawson, M. R. (2008). Connectionism: A hands-on approach. John Wiley & Sons.

Dayan, E., van Keulen, H., Jones, J., Zipori, I., Shmuel, D. and Challa, H. (1993). Development, calibration and validation of a greenhouse tomato growth model: I. description of the model. Agricultural Systems *43*, 145 – 163.

Ehret, D., Lau, A., Bittman, S., Lin, W. and Shelford, T. (2001). Automated monitoring of greenhouse crops. Agronomie *21*, 403–414.

Ehret, D. L., Hill, B. D., Helmer, T. and Edwards, D. R. (2011). Neural network modeling of greenhouse tomato yield, growth and water use from automated crop monitoring data. Computers and electronics in agriculture *79*, 82–89.

Ehret, D. L., Hill, B. D., Raworth, D. A. and Estergaard, B. (2008). Artificial neural network modelling to predict cuticle cracking in greenhouse peppers and tomatoes. computers and electronics in agriculture *61*, 108–116.

Farquhar, G. D. and Sharkey, T. D. (1982). Stomatal conductance and photosynthesis. Annual review of plant physiology *33*, 317–345.

Farquhar, G. D., von Caemmerer, S. and Berry, J. A. (1980). A biochemical model of photosynthetic CO_2 assimilation in leaves of C_3 species. Planta *149*, 78–90.

Farquhar, G. D., von Caemmerer, S. and Berry, J. A. (2001). Models of photosynthesis. Plant Physiology *125*, 42–45.

Fatnassi, H., Boulard, T., Demrati, H., Bouirden, L. and Sappe, G. (2002). SE—structures and environment. Biosystems Engineering *82*, 97 – 105.

Fazlil-Ilahil, W. F. (2009). Evapotranspiration models in greenhouse. PhD thesis, Wageningen Agricultural University, The Netherlands.

Fernández, J. and Cuevas, M. (2010). Irrigation scheduling from stem diameter variations: A review. Agricultural and Forest Meteorology *150*, 135 – 151.

Ferreira, P. M., Faria, E. A. and Ruano, A. E. (2002). Neural network models in greenhouse air temperature prediction. Neurocomputing *43*, 51–75.

Ferreira, P. M., Gomes, J. M., Martins, I. A. C. and Ruano, A. E. (2012). A neural network based intelligent predictive sensor for cloudiness, solar radiation and air temperature. Sensors *12*, 15750–15777.

Fitz-Rodríguez, E. and A. Giacomelli, G. (2009). Yield prediction and growth mode characterization of greenhouse tomatoes with neural networks and fuzzy logic. Transactions of the ASABE *52*, 2115.

Fitz-Rodríguez, E., Kacira, M., Villarreal-Guerrero, F., Giacomelli, G. A., Linker, R., Kubota, C. and Arbel, A. (2012). Neural network predictive control in a naturally ventilated and fog cooled greenhouse. Acta Horticulturae *952*, 45–52.

Fitz-Rodríguez, E., Kubota, C., Giacomelli, G. A., Tignor, M. E., Wilson, S. B. and McMahon, M. (2010). Dynamic modeling and simulation of greenhouse environments under several scenarios: A web-based application. Computers and Electronics in Agriculture *70*, 105 – 116.

Fourati, F. (2014). Multiple neural control of a greenhouse. Neurocomputing *139*, 138 – 144.

Fourati, F. and Chtourou, M. (2007). A greenhouse control with feedforward and recurrent neural networks. Simulation Modelling Practice and Theory *15*, 1016 – 1028.

Franks, P. J., Cowan, I. R. and Farquhar, G. D. (1997). The apparent feedforward response of stomata to air vapour pressure deficit: information revealed by different experimental procedures with two rainforest trees. Plant, Cell & Environment *20*, 142–145.

Fuchs, M., Dayan, E., Shmuel, D. and Zipori, I. (1997). Effects of ventilation on the energy balance of a greenhouse with bare soil. Agricultural and Forest Meteorology *86*, 273–282.

Giacomelli, G., Castilla, N., van Henten, E., Mears, D. and Sase, S. (2007). Innovation in greenhouse engineering. Acta Horticulturae *801*, 75–88.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org.

Green, D. (1996). Cognitive science: An introduction. Wiley-Blackwell.

Guzmán-Cruz, R., Castañeda-Miranda, R., García-Escalante, J. J., López-Cruz, I. L., Lara-Herrera, A. and De la Rosa, J. I. (2009). Calibration of a greenhouse climate model using evolutionary algorithms. Biosystems engineering *104*, 135–142.

Gázquez, J., López, J., Baeza, E., Pérez-Parra, J., Fernández, M., Baille, A. and González-Real, M. (2008). Effects of vapour pressure deficit and radiation on the transpiration rate of a greenhouse sweeet pepper crop. Acta Horticulturae *797*, 259–265.

Haefner, J. W. (1996). Modeling Biological Systems. Springer Nature.

Harmanto, Salokhe, V., Babel, M. and Tantau, H. (2005). Water requirement of drip irrigated tomatoes grown in greenhouse in tropical environment. Agricultural Water Management *71*, 225 – 242.

Hashimoto, Y. (1980). Computer control of short term plant growth by monitoring leaf temperature. Acta Horticulturae *106*, 139–146.

Hashimoto, Y. (1993). Computer Integrated System for the Cultivating Process in Agriculture and Horticulture: Approach to "Intelligent Plant Factory" chapter 8, pp. 175–196. Boston: Academic Press.

He, F. and Ma, C. (2010). Modeling greenhouse air humidity by means of artificial neural network and principal component analysis. Computers and Electronics in Agriculture , *71*, S19–S23.

He, F., Ma, C., Zhang, J. and Chen, Y. (2007). Greenhouse air temperature and humidity prediction based on improved bp neural network and genetic algorithm. In Advances in Neural Networks–ISNN 2007, pp. 973–980. Springer.

Heuvelink, E. (1995). Growth, development and yield of a tomato crop: periodic destructive measurements in a greenhouse. Scientia Horticulturae , *61*, 77 – 99.

Horner, H. and Kühn, R. (1998). Neural Networks, pp. 125–161. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks , *2*, 359 – 366.

Hunt, R. (1990). Basic Growth Analysis. Springer Science + Business Media.

Hunt, S. (2003). Measurements of photosynthesis and respiration in plants. Physiologia Plantarum *117*, 314–325.

Impron, I., Hemming, S. and Bot, G. P. A. (2007). Simple greenhouse climate model as a design tool for greenhouses in tropical lowland. Biosystems Engineering *98*, 79–89.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An introduction to statistical learning. 6 edition, Springer.

Johnson, J. B. and Omland, K. S. (2004). Model selection in ecology and evolution. Trends in Ecology & Evolution *19*, 101 – 108.

Jolliet, O. and Bailey, B. J. (1992). The effect of climate on tomato transpiration in greenhouses: measurements and models comparison. Agricultural and Forest Meteorology *58*, 43–62.

Jones, H. (1998). Stomatal control of photosynthesis and transpiration. Journal of Experimental Botany *49*, 387–398.

Jones, H. and Tardieu, F. (1998). Modelling water relations of horticultural crops: a review. Scientia Horticulturae *74*, 21 – 46.

Jones, H. G. (2004). Irrigation scheduling: advantages and pitfalls of plant-based methods. Journal of Experimental Botany *55*, 2427.

Jones, H. G. (2014). Plants and Microclimate. Cambridge University Press.

Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., Godfray, H. C. J., Herrero, M., Howitt, R. E., Janssen, S., Keating, B. A., Munoz-Carpena, R., Porter, C. H., Rosenzweig, C. and Wheeler, T. R. (2016). Brief history of agricultural systems modeling. Agricultural Systems *Article in press*, –.

Katsoulas, N., Baille, A. and Kittas, C. (2001). Effect of misting on transpiration and conductances of a greenhouse rose canopy. Agricultural and forest meteorology *106*, 233–247.

Katsoulas, N., Baille, A. and Kittas, C. (2002). SE—structures and environment: Influence of leaf area index on canopy energy partitioning and greenhouse cooling requirements. Biosystems engineering *83*, 349–359.

Katsoulas, N., Baille, A. and Kittas, C. (2007). Leaf boundary layer conductance in ventilated greenhouses: An experimental approach. Agricultural and Forest Meteorology *144*, 180 – 192.

Kawamitsu, Y., Yoda, S. and Agata, W. (1993). Humidity pretreatment affects the responses of stomata and CO_2 assimilation to vapor pressure difference in C_3 and C_4 plants. Plant and Cell Physiology *34*, 113.

Kimball, B. A. (1973). Simulation of the energy balance of a greenhouse. Agricultural meteorology *11*, 243–260.

Kittas, C. and Bartzanas, T. (2007). Greenhouse microclimate and dehumidification effectiveness under different ventilator configurations. Building and Environment *42*, 3774 – 3784.

Kittas, C., Katsoulas, N. and Baille, A. (1999). Transpiration and canopy resistance of greenhouse soisoill roses: Measurements and modeling. Acta Horticulturae *507*, 61–68.

Kläring, H.-P., Hauschild, C., Heißner, A. and Bar-Yosef, B. (2007). Model-based control of CO_2 concentration in greenhouses at ambient levels increases cucumber yield. Agricultural and forest meteorology *143*, 208–216.

Kläring, H.-P. and Krumbein, A. (2013). The effect of constraining the intensity of solar radiation on the photosynthesis, growth, yield and product quality of tomato. Journal of Agronomy and Crop Sience *199*.

Kramer, P. J. and Boyer, J. S. (1995). Water Relations of Plants and Soils. Academic Press.

Kuo, B. C. (1995). Automatic control systems. 7 edition, John Wiley & Sons Inc.

Lentz, W. (1998). Model applications in horticulture: a review. Scientia horticulturae *74*, 151–174.

Lin, W. and Hill, B. (2008). Neural network modelling to predict weekly yields of sweet peppers in a commercial greenhouse. Canadian Journal of Plant Science *88*, 531–536.

Linker, R., Gutman, P. and Seginer, I. (2000). Robust model-based failure detection and identification in greenhouses. Computers and Electronics in Agriculture *26*, 255–270.

Linker, R. and Seginer, I. (2004). Greenhouse temperature modeling: a comparison between sigmoid neural networks and hybrid models. Mathematics and computers in simulation *65*, 19–29.

Linker, R., Seginer, I. and Gutman, P. (1998). Optimal CO_2 control in a greenhouse modeled with neural networks. Computers and Electronics in Agriculture *19*, 289 – 310.

Ljung, L. (2010). Perspectives on system identification. Annual Reviews in Control *34*, 1–12.

López-Cruz, I. L., Fitz-Rodríguez, E., Torres-Monsivais, J. C., Trejo-Zúñiga, E. C., Ruíz-García, A. and Ramírez-Arias, A. (2014). Control Strategies of Greenhouse Climate for Vegetables Production pp. 401–421. Cham: Springer International Publishing.

López-Cruz, I., Ramírez-Arias, A., Rojano-Aguilar, A. and Ruiz-García, A. (2008). Modeling of greenhouse climate using evolutionary algorithms. Acta Horticulturae , *801*, 401–408.

López-Cruz, I., Rojano-Aguilar, A., Ojeda-Bustamante, W. and Salazar-Moreno, R. (2007). ARX models for predicting greenhouse air temp. Agrociencia , *41*, 181–192.

Maclean, H., Dochain, D., Waters, G., Stasiak, M., Dixon, M. and Van Der Straeten, D. (2012). A model development approach to ensure identifiability of a simple mass balance model for photosynthesis and respiration in a plant growth chamber. Ecological Modelling , *246*, 105–118.

Marcelis, L., Heuvelink, E. and Goudriaan, J. (1998). Modelling biomass production and yield of horticultural crops: a review. Scientia Horticulturae , *74*, 83 – 111.

Masters, T. (1993). Practical neural network recipes in C++. Morgan Kaufmann.

Medrano, E., Lorenzo, P., Sánchez-Guerrero, M. C. and Montero, J. I. (2005). Evaluation and modelling of greenhouse cucumber-crop transpiration under high and low radiation conditions. Scientia Horticulturae *105*, 163–175.

Menezes, J. M. P. and Barreto, G. A. (2008). Long-term time series prediction with the NARX network: an empirical evaluation. Neurocomputing *71*, 3335–3343.

Menzel, W. (1998). Problem Solving with Neural Networks pp. 162–177. Berlin, Heidelberg: Springer Berlin Heidelberg.

Millan-Almaraz, J. R., Guevara-Gonzalez, R. G., Romero-Troncoso, R., Osornio-Rios, R. A. and Torres-Pacheco, I. (2009). Advantages and disadvantages on photosynthesis measurement techniques: A review. African Journal of Biotechnology , *8*.

Monsi, M. and Saeki, T. (1953). Über den lichtfaktor in den pflanzengesellschaften und seine bedeutung für die stoffproduktion. Japanese Journal of Botany , *14*, 22–52.

Monsi, M. and Saeki, T. (2005). On the factor light in plant communities and its importance for matter production. Annals of Botany , *95*, 549–567.

Monteith, J. and Unsworth, M. (2013). Principles of environmental physics: plants, animals, and the atmosphere. 4 edition, Academic Press.

Monteith, J. L. (1995). A reinterpretation of stomatal responses to humidity. Plant, Cell & Environment *18*, 357–364.

Morimoto, T. and Hashimoto, Y. (1996). Optimal control of plant growth in hydroponics using neural networks and genetic algorithms. Acta Horticulturae *406*, 433–440.

Murata, N., Yoshizawa, S. and Amari, S.-i. (1994). Network information criterion - determining the number of hidden units for an artificial neural network model. IEEE Transactions on Neural Networks *5*, 865–872.

Nederhoff, E. (1995). Carbon dioxide balance. In Greenhouse Climate Control - an integrated approach, (Bakker, J. C., Bot, G. P. A., Challa, H. and van de Braak, N. J., eds), pp. 151–155. Wageningen Press.

Nguyen, D. and Widrow, B. (1990a). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In 1990 IJCNN International Joint Conference on Neural Networks pp. 21–26 vol.3, IEEE.

Nguyen, D. H. and Widrow, B. (1990b). Neural networks for self-learning control systems. Control Systems Magazine, IEEE *10*, 18–23.

Nilsson, N. J. (2009). The Quest for Artificial Intelligence. Cambridge University Press.

Ogata, K. (2002). Modern Control Engineering. 4 edition, Pearson.

Oussar, Y. and Dreyfus, G. (2001). How to be a gray box: dynamic semi-physical modeling. Neural Networks *14*, 1161–1172.

Pasgianos, G., Arvanitis, K., Polycarpou, P. and Sigrimis, N. (2003). A nonlinear feedback technique for greenhouse environmental control. Computers and Electronics in Agriculture *40*, 153 – 177.

Patil, S., Tantau, H. and Salokhe, V. (2008). Modelling of tropical greenhouse temperature by auto regressive and neural network models. Biosystems Engineering *99*, 423 – 431.

Pearcy, R. W., Osteryoung, K. and Calkin, H. W. (1985). Photosynthetic responses to dynamic light environments by hawaiian trees: Time course of CO_2 uptake and carbon gain during sunflecks. Plant Physiology *79*, 896–902.

Pearl, J. (2009). Causality: Models, Reasoning and Inference. Cambridge University Press.

Pecher, D. and Zwaan, R. A. (2005). Grounding cognition: The role of perception and action in memory, language, and thinking. Cambridge University Press.

Pfeifer, R. and Bongard, J. (2006). How the body shapes the way we think: a new view of intelligence. MIT press.

Pfeifer, R. and Iida, F. (2004). Embodied artificial intelligence: Trends and challenges. In Embodied artificial intelligence pp. 1–26. Springer.

Pfeifer, R. and Scheier, C. (2001). Understanding intelligence. MIT press.

Riedmiller, M. (1994). Advanced supervised learning in multi-layer perceptrons — from backpropagation to adaptive learning algorithms. Computer Standards & Interfaces *16*, 265 – 278.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the rprop algorithm. In International Conference on Neural Networks pp. 586–591 vol.1, IEEE.

Ripley, B. D. (2006). Pattern recognition and neural networks. Cambridge university press.

Rouphael, Y. and Colla, G. (2005). Radiation and water use efficiencies of greenhouse zucchini squash in relation to different climate parameters. European Journal of Agronomy *23*, 183–194.

Roy, J. C., Boulard, T., Kittas, C. and Wang, S. (2002). Convective and ventilation transfers in greenhouses, part 1: the greenhouse considered as a perfectly stirred tank. Biosystems Engineering *83*, 1–20.

Russell, S. J. and Norvig, P. (2010). Artificial Intelligence: A Modern Approach. Pearson.

Salazar, R., López, I., Rojano, A., Schmidt, U. and Dannehl, D. (2015). Tomato yield prediction in a semi-closed greenhouse. Acta Horticulturae *1107*, 263–270.

Salazar, R., Rojano, A., Lopez, I. and Schmidt, U. (2010a). A model for the combine description of the temperature and relative humidity

regime in the greenhouse. In 2010 Ninth Mexican International Conference on Artificial Intelligence pp. 113–117, Institute of Electrical and Electronics Engineers (IEEE).

Salazar, R., Schmidt, U., Huber, C., Rojano, A. and Lopez, I. (2010b). Neural networks models for temperature and CO_2 control. International Journal of Agricultural Research *5*, 191–200.

Salazar-Moreno, R., López-Cruz, I. and Rojano-Aguilar, A. (2008). A neural network model to predict temperature and relative humidity in a greenhouse. Acta Horticulturae *801*, 539–545.

Salazar-Moreno, R., Rojano-Aguilar, A., Schmidt, U. and Huber, C. (2011). Temperature and co2 prediction to control greenhouse environment. Acta Horticulturae *893*, 689–696.

Saxén, H. and Pettersson, F. (2006). Method for the selection of inputs and structure of feedforward neural networks. Computers & chemical engineering *30*, 1038–1045.

Schillaci, G., Hafner, V. V. and Lara, B. (2012). Coupled inverse-forward models for action execution leading to tool-use in a humanoid robot. In Human-Robot Interaction, Seventh annual ACM/IEEE international conference on pp. 231–232, ACM.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks *61*, 85 – 117.

Schmidt, U. (1992). On-line aqcuisition of transpiration as a basis for ecological process control in greenhouses. Acta Horticulturae *304*, 353–362.

Schmidt, U. (1996). Greenhouse climate control with a combine model of greenhouse and plant by using online measurement of leaf temperature and transpiration. Acta Horticulturae *406*, 89–98.

Schmidt, U. (2002). Modelling of stomatal conductance as a variable for environmental control in greenhouses. Acta Horticulturae *593*, 227–234.

Schmidt, U. (2005). Microclimate control in greenhouses based on phytomonitoring data and mollier phase diagram. Acta Horticulturae *691*, 125–132.

Schmidt, U. and Exarchou, E. (2000). Controlling of irrigation systems of greenhouse plants by using measured transpiration sum. Acta Horticulturae *537*, 487–494.

Schmidt, U., Huber, C., Rocksch, T., Salazar-Moreno, R. and Rojano-Aguilar, A. (2008). Greenhouse cooling and carbon dioxide fixation by using high pressure fog systems and phytocontrol strategy. Acta Horticulturae *797*, 279–284.

Schmidt, U., Schuch, I., Dannehl, D., Rocksch, T., Salazar-Moreno, R., Rojano-Aguilar, A. and López-Cruz, I. (2014). Long time plant response measurements for yield prediction, water use and climate control optimization using gas exchange measurements in semi closed and ventilated greenhouses. Acta Horticulturae *1037*, 469–476.

Scholberg, J., McNeal, B. L., Jones, J. W., Boote, K. J., Stanley, C. D. and Obreza, T. A. (2000). Growth and canopy characteristics of field-grown tomato. Agronomy Journal *92*, 152–159.

Seginer, I. (1997). Some artificial neural network applications to greenhouse environmental control. Computers and Electronics in Agriculture *18*, 167–186.

Seginer, I. (2002). SE—structures and environment: The penman–monteith evapotranspiration equation as an element in greenhouse ventilation design. Biosystems Engineering *82*, 423 – 439.

Seginer, I., Boulard, T. and Bailey, B. (1994). Neural network models of the greenhouse climate. Journal of Agricultural Engineering Research *59*, 203–216.

Seginer, I., Hwang, Y., Boulard, T. and Jones, J. W. (1996). Mimicking an expert greenhouse grower with a neural-net policy. Transactions of the ASAE *39*, 299–306.

Seginer, I. and Kantz, D. (1986). *In-situ* determination of transfer coefficients for heat and water vapour in a small greenhouse. Journal of agricultural engineering research *35*, 39–54.

Seginer, I. and Sher, A. (1993). Optimal greenhouse temperature trajectories for a multi-state-variable tomato model. In The Computerized Greenhouse pp. 153–172. Elsevier.

Shimizu, H., Moriizumi, S. and Wada, M. (2004). Soft sensing of plant shoot-tip temperature. In SICE 2004 Annual Conference vol. 2, pp. 1011–1014 vol. 2, IEEE Society of Instrument and Control Engineers.

Stanghellini, C. (1987). Transpiration of greenhouse crops : an aid to climate management. PhD thesis, Instituut voor Mechanisatie, Arbeid en Gebouwen (IMAG) Wageningen, The Netherlands.

Stanghellini, C. (1988). Microclimate and transpiration of greenhouse crops. Acta Horticulturae *229*, 405–414.

Stanghellini, C. and de Jong, T. (1995). A model of humidity and its applications in a greenhouse. Agricultural and Forest Meteorology *76*, 129–148.

Stathakis, D. (2009). How many hidden layers and nodes? International Journal of Remote Sensing *30*, 2133–2147.

Steduto, P., Hsiao, T. C., Fereres, E. and Raes, D. (2012). Crop yield response to water. FAO Irrigation and drainage paper *66*.

Steppe, K., De Pauw, D. J. and Lemeur, R. (2008). A step towards new irrigation scheduling strategies using plant-based measurements and mathematical modelling. Irrigation Science *26*, 505–517.

Sánchez, J. A., Rodríguez, F., Guzmán, J. L. and Arahal, M. R. (2012). Virtual sensors for designing irrigation controllers in greenhouses. Sensors *12*, 15244–15266.

Takakura, T. (1993). Climate under Cover. Springer Netherlands.

Takakura, T. and Fang, W. (2002). Climate under cover. Kluwer Academic Publishers.

Takakura, T., Goudriaan, J. and Louwerse, W. (1975). A behaviour model to simulate stomatal resistance. Agricultural Meteorology *15*, 393 – 404.

Takakura, T., Kubota, C., Sase, S., Hayashi, M., Ishii, M., Takayama, K., Nishina, H., Kurata, K. and Giacomelli, G. (2009). Measurement of evapotranspiration rate in a single-span greenhouse using the energy-balance equation. Biosystems Engineering *102*, 298 – 304.

Takakura, T., Sunagawa, H., Tamaki, M., Usui, T. and Taniai, N. (2017). In site net photosynthesis measurement of a plant canopy in a single-span greenhouse. Journal of Advances in Agriculture *7*, 1015–1020.

Takakura, T., Takayama, K., Nishina, H., Tamura, K. and Muta, S. (2005). Evapotranspiration estimate by heat balance equation. In 2005 ASAE Annual Meeting p. 1, American Society of Agricultural Engineers.

Talbott, L. D., Rahveh, E. and Zeiger, E. (2003). Relative humidity is a key factor in the acclimation of the stomatal response to CO_2. Journal of Experimental Botany *54*, 2141.

Tantau, H.-J. (1989). Models for greenhouse climate control. Acta Horticulturae *245*, 397–405.

Thornley, J. H. M. and France, J. (2007). Mathematical models in agriculture: quantitative methods for the plant, animal and ecological sciences. Cabi.

Ton, Y., Kopyt, M. and Nilov, N. (2004). Phytomonitoring technique for tuning irrigation of vineyards. Acta Horticulturae *646*, 133–142.

Trejo-Perea, M., Herrera-Ruiz, G., Ríos-Moreno, J., Castañeda-Miranda, R. and Rivas-Araiza, E. (2009). Greenhouse energy consumption prediction using neural networks models. International Journal of Agriculture & Biology *11*, 1–6.

Tsafaras, I. and de Koning, A. (2017). Real-time application of crop transpiration and photosynthesis models in greenhouse process control. Acta Horticulturae *1154*, 65–72.

Uchida-Frausto, H., Pieters, J. and Deltour, J. (2003). Modelling greenhouse temperature by means of auto regressive models. Biosystems Engineering *84*, 147–157.

Uchida-Frausto, H. and Pieters, J. G. (2004). Modelling greenhouse temperature using system identification by means of neural networks. Neurocomputing *56*, 423–428.

Udink ten Cate, A., Bot, G. and van Dixhoorn, J. (1978). Computer control of greenhouse climates. Acta Horticulturae *87*, 265–272.

Udink ten Cate, A. and Challa, H. (1984). On optimal computer control of the crop growth system. Acta Horticulturae *148*, 267–276.

Udink ten Cate, A. J. (1985). Modelling and simulation in greenhouse climate control. Acta Horticulturae *174*, 461–468.

Van Henten, E. J. (2003). Sensitivity analysis of an optimal control problem in greenhouse climate management. Biosystems Engineering *85*, 355–364.

Van Straten, G. and Van Henten, E. (2010). Optimal greenhouse cultivation control: survey and perspectives. IFAC Proceedings Volumes *43*, 18–33.

Villarreal-Guerrero, F., Kacira, M., Fitz-Rodriguez, E., Kubota, C., Giacomelli, G. A., Linker, R. and Arbel, A. (2012a). Comparison of three evapotranspiration models for a greenhouse cooling strategy with natural ventilation and variable high pressure fogging. Scientia Horticulturae *134*, 210–221.

Villarreal-Guerrero, F., Kacira, M., Fitz-Rodríguez, E., Linker, R., Kubota, C., Giacomelli, G. A. and Arbel, A. (2012b). Simulated performance of a greenhouse cooling control strategy with natural ventilation and fog cooling. Biosystems Engineering *111*, 217–228.

Von Caemmerer, S. v. and Farquhar, G. (1981). Some relationships between the biochemistry of photosynthesis and the gas exchange of leaves. Planta *153*, 376–387.

von Zabeltitz, C. (2011). Integrated Greenhouse Systems for Mild Climates. Springer.

Wallach, D., Makowski, D., Jones, J. W. and Brun, F. (2014). Working with Dynamic Crop Models. Second edition edition, Elsevier, San Diego.

Wang, S., Boulard, T. and Haxaire, R. (1999). Air speed profiles in a naturally ventilated greenhouse with a tomato crop. Agricultural and Forest Meteorology *96*, 181 – 188.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. Neural Networks *1*, 339–356.

Wong, S. C., Cowan, I. R. and Farquhar, G. D. (1979). Stomatal conductance correlates with photosynthetic capacity. Nature *282*, 424–426.

Zee, F. and Bubenheim, D. (1997). Plant growth model using artifical neural networks. In SAE Technical Paper SAE International.

Ziemke, T., Jirenhed, D.-A. and Hesslow, G. (2005). Internal simulation of perception: a minimal neuro-robotic model. Neurocomputing *68*, 85–104.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständing und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe.

Berlin, den 15. November 2017

Luis Carlos Miranda Trujillo